University of Alberta

Proteomics Tools for Drug Discovery

By

Zhan Chang

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of

Master of Science

In

Pharmaceutical Sciences

Faculty of
Pharmacy and Pharmaceutical Sciences
Edmonton, Alberta
Fall, 2003

National Library
of Canada

Acquisitions and
Bibliographic Services

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale
du Canada

Acquisisitons et
services bibliographiques

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 0-612-87733-7
Our file Notre référence
ISBN: 0-612-87733-7

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou aturement reproduits sans son autorisation.

Canada

University of Alberta

Library Release Form

Name of Author: Zhan Chang

Title of Thesis: Proteomics Tools for Drug Discovery

Degree: Master of Science

Year This Degree Granted: 2003

Permission is hereby granted to the University of Alberta Library to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves all other publication and other rights in association with the copyright in the thesis, and except as herein before provided, neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatever without the author's prior written permission.

Oct. 3 / 2003

University of Alberta

Faculty of Graduate Studies and Research

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled "Proteomics Tools for Drugs Discovery" by Zhan Chang in partial fulfillment of the requirements for the degree of Master of Science in Pharmaceutical Sciences.

Date: Oct.

Abstract

Given the growing importance of proteomics in life sciences, it is little wonder that proteomics is also becoming integral to many pharmaceutical research projects. This thesis is about the development and application of new proteomics tools towards drug discovery and development. In Chapter 2, a web-based gel management system, called Gelscape is described. Gelscape facilitates analysis and archiving of 1D and 2D protein gels using a variety of web tools and web scripts. In addition to Gelscape, another open access web-tool called DrugBank for proteomic based drug discovery is described in Chapter 3. DrugBank is a web-enabled database that combines drug information with drug target information to allow users the possibility of linking small molecule data with protein sequence/structure data. In Chapter 4, we demonstrate how experimental methods in proteomics can be integrated with computational tools to characterize a protein (from *Methanobacterium thermoautotrophicum* Δh) with unknown function and structure.

Acknowledgements

I would like to give my whole-hearted thanks to Dr. David Wishart, for his patience, support and guidance. Also, I would like thank Haiyan Zhang, Lena Andrew, Godwin Amegbey and the rest of Dr. Wishart's lab for the assistance and friendship they have given me over the past 2 years. Finally, I would like to give my thanks to the Faculty of Pharmacy and Pharmaceutical Sciences, as well as the Protein Engineering Network of Centre of Excellence (PENCE) for financial support.

Table of Contents

1. Introduction 1

1.1. Introduction 1

1.2. Proteomics 2

1.3. Proteomics in Pharmaceutical Research 4

1.4. Two-Dimensional Gels and Proteomics 7

1.5. Bioinformatics and Drug Discovery 13

1.6. Drug Databases 18

1.7. Structural Proteomics 21

1.8. Outline of This Dissertation 25

1.9. My Contribution to This Dissertation 26

2. GelScape 28

2.1. Introduction 28

2.2. System and Methods 30

2.2.1. Platform and Availability 30

2.2.2. Data Sources, Storage, Retrieval and Display 31

2.3. Implementation 35

2.3.1. General GelScape Interface and Home Page 35

2.3.2. GelScape Image Maps 37

2.3.3. Uploading and Comparing Gel Images 38

2.3.4. Applications 39

2.4. Discussion 42

2.4.1. Database Construction and Information Display 42

2.4.2. Database Administration and Maintenance 43

2.4.3. Future Directions 44

2.5. Conclusions 45

3. DrugBank 46

3.1. Introduction 46

3.2. System and Methods 48

3.2.1. Platform, Software and Availability 48

3.2.2. Data Sources and Storage 49

3.2.3. DrugBank Structure and Layout 52

3.2.4. Data Retrieval and Searching 56

3.2.5. The Data Extractor 57

3.2.6. Structural Query 59

3.2.7. Other Search Utilities 60

3.3. Discussion 63

3.3.1. Data Integration 63

3.3.2. Database Administration and Maintenance 64

3.4. Conclusions and Future Directions 65

4. Expression, Purification and Preliminary Characterization of an Unknown Protein - M.t.0776 68

4.1. Introduction 68

4.1.1. *Methanobacterium thermoautotrophicum* (M.t.) 68

4.1.2. Affinity Chromatography in Protein Purification 72

4.2. Materials and Methods 73
4.2.1. Materials 73
4.2.2. Transformation 75
4.2.3. Expression 77
4.2.4. Purification 78
4.2.5. Mass Spectrometric Analysis of the Gene Product 78
4.2.6. NMR Sample Preparation and Spectra Collection 79
4.2.7. Bioinformatic Analysis 80
4.3. Results and Discussion 81
4.3.1. Transformation, Expression and Purification 81
4.3.2. MS and NMR Analysis 85
4.3.3. Bioinformatic Analysis of M.t. 0776 Protein 88
4.3.3.1. Secondary Structure 89
4.3.3.2. Hydrophobicity Distribution 91
4.3.3.3. Preliminary 3D Structure Prediction 92
4.4. Conclusion 96
5. General Discussion and Conclusions 98
References 103

List of Tables

Table 1.1 The New Paradigm in Drug Discovery and Design.....16

List of Figures

Figure 1.1 An Example of a High Resolution 2D gel, Acquired by the Max Delbrück Center 9

Figure 1.2 A Schematic Illustration of the Two-dimensional Electrophoresis Process 11

Figure 1.3 The Exponential Growth in the Number of the Total Sequences Deposited in the GenBank 14

Figure 1.4 FDA Generic Drug Approval Rates for The Past 7 Years. 19

Figure 1.5 Procedures for Comparative Modeling 22

Figure 2.1(a) The Layout of GelScape 33

Figure 2.1(b) The Layout of GelScape 34

Figure 2.2 Overview of the GelScape Implementation 36

Figure 2.3 Annotated 2D Gels of Human Heart Atrium (A) and Human Heart Ventricle (B)..... 41

Figure 3.1 The DrugBank Homepage and the Drug List 50

Figure 3.2 A Screenshot of a Sample DrugCard in DrugBank 54

Figure 3.3 A Screen Shot of DrugBank's Structural Query Function 58

Figure 4.1 The Genome and Protein Sequence of M.t.0776 70

Figure 4.2 The BLAST Search Result of M.t.0776 by NCBI BLASTP 2.2.5 Conducted on September 1, 2003 71

Figure 4.3 Interactions between Neighboring Residues in the 6xHis Tag and Ni-NTA Matrix 74

Figure 4.4 Construction of the M.t.0776 Expression Vector 76

Figure 4.5 1D SDS-PAGE Gel of the M.t. 0776 Expression Levels..... 83

Figure 4.6 SDS-PAGE Gel of the Purification Result of M.t.0776 84

Figure 4.7 Mass Spectroscopy Analysis by IBD of An SDS-PAGE Gel in the Purification of M.t. 0776 86

Figure 4.8 1D ^{1}H NMR Trace of M.t. 0776 Acquired at 500 MHz 87

Figure 4.9 Prediction of M.t. 0776's Secondary Structure by NNPREDICT .. 90

Figure 4.10 Hydrophobicity Distribution Analysis of M.t. 0776 by ProtScale 93

Figure 4.11 Possible Structures of M.t. 0776 Predicted by 3D-PSSM 94

Figure 4.12 A 2D Structure Alignment of M.t. 0776 (M) as generated by NNPREDICT and the Ferredoxin Like Metal Binding Domain of ATX1 Metallochaperone Protein, 1CC8 (F) 95

List of Abbreviations

2-DE (2-Dimensional Electrophoresis)

BLAST (Basic Local Alignment Search Tool)

CAS (Chemistry Abstract Service)

CGI (Common Gateway Interface)

DIN (Drug Identification Number)

ExPASy (Expert Protein Analysis System)

HTML (Hypertext Markup Language)

IEF (Isoelectric Focusing)

IPG (Immobilized pH Gradients)

IPTG (isopropyl-beta-D-thiogalactopyranoside)

JPEG (Joint Photographic Experts Group)

M.t. (*Methanobacterium thermoautotrophicum*)

MS (Mass Spectroscopy)

MSDS (Material Safety Data Sheet)

NMR (Nuclear Magnetic Resonance)

OCI (Ontario Cancer Institute)

PDB (Protein Data Bank)

PERL (Practical Extraction and Report Language)

PIR (Protein Information Resource)

SDS-PAGE (Sodium Dodecyl Sulfate Polyacrylamide Gel Electrophoresis)

SMILES (Simplified Molecular Input Line Entry System)

SNP (Single nucleotide polymorphism)

HPLC (High Pressure Liquid Chromatography)

Chapter 1; Introduction

1.1 Introduction

This thesis is concerned with the development and testing of novel techniques in proteomics research. As such, the research described here is primarily focused on technique or technology development as opposed to the more conventional hypothesis driven research. To gain a broad level of experience and understanding, I have chosen three areas of proteomics research and have attempted to address or partially address several outstanding problems in these areas by using a combination of bioinformatics, computer programming and experimental protein chemistry. The first area of proteomics research concerns the analysis, storage and manipulation of 2D gel images. A continuing problem in 2D gel analysis is the cost and difficulty of storing, viewing and sharing this data between laboratories. In Chapter 2 of this thesis, I describe the development of a novel web-based gel viewing and archiving package that addresses these problems through the use of dynamic HTML and Java Applets. The second area of proteomics research I chose to explore concerns the linkage of protein sequence/structure data to small molecule drug data. A major problem in this area is that this information is highly dispersed and not readily available *via* electronic means. In Chapter 3 of this thesis, I describe the development of a unique, web-based databank called DrugBank. This represents one of the first and perhaps most ambitious efforts to electronically link drug information to drug target (i.e.

proteomics) information. The final area of proteomics research I chose to investigate concerns structural proteomics. In Chapter 4 of this thesis, I describe my efforts to express, purify and structurally characterize (spectroscopically and computationally – *via* bioinformatics) a 101 residue protein of unknown function and structure (M.t. 0776) isolated from the archeon *Methanobacterium thermoautotrophicum*. Collectively, these efforts represent a relatively modest contribution to proteomics research, but I believe they could serve as the basis for more far reaching developments in the field of structural and functional proteomics.

1.2 Proteomics

While some have called the final decade of the 20th century the "Decade of Genomics", many others have labeled the first decade of the 21st century the "Decade of Proteomics" (Anderson et al., 2000). Whether this decade will live up to its name remains to be seen, but there can be little doubt that proteomics is a very "hot" area of biomedical research. Proteomics is derived from the word proteome which refers to the proteins encoded by a genome. Specifically, proteomics can be defined as the large scale study of proteins *via* biochemical methods (Pandey and Mann, 2000). The "Decade of Proteomics" epithet serves to emphasize that the major challenge of biomedical research today involves the characterization of the properties and biological functions of not only genes, but proteins. It is through proteomics research that we can learn how proteins and

peptides are involved in human disease, how they can be developed into new therapies and how they can be used for other industrial or pharmaceutical applications. Because of its shear complexity, proteomics research will require knowledge from almost all areas of biological science, including genetics, developmental biology, cell and molecular biology, physiology, protein chemistry, as well as structural biology.

Nowadays proteomics can be divided into two broad categories: 1) expressional proteomics and 2) functional proteomics. Expressional proteomics is concerned with the identification and measurement of protein expression levels in complete proteomes, whereas functional proteomics is concerned with determining the function, interactions and structures of more limited protein sets. Expressional proteomic analyses are usually carried out using two-dimensional gel electrophoresis (2-DE) to facilitate protein separation followed by protein identification by mass spectrometry (MS) and computerized database searches (Sagliocco et al., 1998; Shevchenko et al., 1996). Functional proteomics employs a broader range of techniques including 2-DE, MS, 2D HPLC, yeast 2-hybrid analysis (Uhrig et al., 1999), affinity pull-down experiments (Magga et al., 2000), X-ray crystallography and NMR spectroscopy. Both branches of proteomics are rapidly expanding and the technologies for both are rapidly evolving.

1.3 Proteomics in Pharmaceutical Research

Proteomics promises to help in at least two key areas of pharmaceutical research. First, it will allow researchers to identify or work with many more potential drug targets (i.e. proteins). Second, it will allow drug leads to be screened and characterized far more efficiently. To understand how this could come about, consider the following data: By several estimates, essentially all the drugs available today work on only ~500 unique protein targets (Dean and Zanders, 2002; Maggio and Ramnarayan, 2001). This is a remarkably small number when we realize that the human genome project identified ~30 000 unique genes (Venter et al., 2001) – any one of which could be a potential new drug target. Alternate splicing, in conjuction with post-translational modification, suggests that the true size of the human proteome is not 30,000, but more likely 300,000 proteins. Among pharmacological model species, such as the dog and rat, and even microbial target species, such as *S. aureus* and *E. coli* (hamburger disease), there are tens of thousands of genes – and therefore proteins – of potential interest to pharmaceutical scientists. Combining the number of total proteins from human and other species (10^6) with the estimated 10^{200} potential small molecules (Maggio and Ramnarayan, 2001) accessible through combinatorial chemistry, it should make one thing clear: numerous drug targets and their corresponding drugs will no doubt emerge from proteomic research conducted over the next decade.

While the possibilities of 1000's of new targets are causes for considerable excitement, this has to be tempered by the fact that with today's technology, drug development remains costly, time-consuming and prone to failure. Indeed, the drug development process - from initial identification of a disease-associated protein to the market launch of a viable drug can take up to ten years and cost in excess of $800,000,000 (Kara Bio annual report, 2000). For every drug candidate that enters human clinical trials, up to 1000 other candidates will have failed at various stages of discovery and testing. For every drug that successfully emerges from phase III clinical trials, as many as ten others will have failed at the phase III level. According to industry and government data (http://www.bls.gov/oco/cg/cgs009.htm), only one in every 5,000 to 10,000 screened compounds eventually becomes an approved drug. Furthermore, only three out of 10 approved drugs achieve sufficient sales to recover average research and development costs. One consequence of this low success rate is that pharmaceutical researchers have a tendency to pursue lower risk projects. For instance, they tend to focus on developing drug candidates that act on proteins that are already targets of successful drugs. While this approach can increase the probability of discovering improvements over existing drugs, new drugs that work on established targets do not act in fundamentally novel ways. A major challenge for drug development therefore is to increase both the number and range of potential protein targets from approximately 500 today to thousands of targets tomorrow.

The primary reasons for drug candidates failing during development and testing are 1) lack of efficacy 2) unacceptable side effects. To help address these problems, a growing number of pharmaceutical companies are using proteomics tools to more rapidly and efficiently validate new targets for drug development. Proteomics techniques are also being used to refine the selection of compounds that act on identified and validated targets, so that compounds that lack promise, either for efficacy or safety reasons, can be eliminated from development at an early stage. A good example of how proteomics facilitates the drug development process can be seen in an early study of the toxic activity of cyclosporine A (CsA) in kidneys (Sandra et al., 1996). In this study, 2D gel profiles of rat kidney proteins with or without treatment with CsA were compared. One of the down-regulated protein spots arising from CsA treatment was identified as calbindin. Calbindin is a small protein found in kidney tubules which is involved in calcium binding and transport. This study showed that the toxic effects of CsA with respect to intratubular calcification are linked to a decrease in intracellular concentration of calbindin. In essence, proteomics *via* 2D-gel analysis of kidney tissue provided key insights into the understanding of the toxic side effects of CsA. A large number of proteomics studies have also been published showing the application of proteomics in the early detection of cancer (Srinivas et al., 2001; Brichory et al., 2001), in cancer biomarker discovery (Srinivas et al., 2002), in the identification of regulated targets in signal transduction pathways (Lewis et al., 2000) and in the identification of proteases most suitable for drug targeting

(McKerrow et al., 2000). In many of these examples, two-dimensional electrophoresis (2-DE) was a critical part of the discovery process.

1.4 Two-Dimensional Gels and Proteomics

Two-dimensional gel electrophoresis (2-DE) is among the many core technologies in proteomics. Using this method, researchers can separate complex protein mixtures with a remarkably high degree of resolution. The results of the separation in conjunction with mass spectrometric analysis followed by peptide mass database searches is one of the most popular and effective methods for protein identification (Clauser et al., 1995; Fernandez et al., 1998). 2-DE was first introduced in the mid-70's (MacGillivray, 1974; Klose, 1975; O'Farrell, 1975). Simply stated, 2-DE is an electrophoretic separation technique that separates proteins according to charge (in one dimension) and molecular weight (in the second dimension). The resolution achievable by 2-DE makes it possible to analyze hundreds or even thousands of gene products in complex protein mixtures (Figure 1.1). The introduction of immobilized pH gradients (IPG) strips in particular has played a key role to help standardize 2-DE protocol and 2-DE results. It not only simplified the experimental process, but also greatly improved the reproducibility of 2-DE (Bjellqvist et al., 1982; Görg et al., 2000). With the development of microanalytical techniques to identify small amounts of proteins *via* Edman sequencing (Matsudaira, 1987; Aebersold, 1987) or mass spectrometry

(Yates, 1993; James, 1994), 2-DE has gained considerable popularity in proteomics research (Celis, 1995; Rabilloud, 1998; Langen, 2000).

To separate proteins on the basis of charge, 2-DE exploits a technique called isoelectric focusing (IEF). In IEF, proteins are separated on the basis of their intrinsically different isoelectric points (pI) or the pH at which the net charge of the protein is zero. Depending on the local pH of the surrounding solvent, proteins can carry either a net positive, or negative charge because of the presence of differently charged amino acids (Lys, Arg, His are positive; Asp, Glu are negative). When a protein's net charge is zero, it cannot migrate in an electric field. Therefore if a protein is placed in a local solution with a pH gradient and is subjected to an electric field, it will move towards the electrode with the opposite charge (Figure 1.2). At the same time it is moving, it will also pick up or lose protons. As it is picking up or losing protons, it will slow down as its net charge approaches zero. Eventually, the protein will stop at the point where the pH is equal to its pI, because at this point it is electrically neutral. In this way, the protein is "focused" into a single spot or band. In IEF, a pH gradient is normally set up inside the gel using a mixture of ampholytes (zwitter-ionic small molecules) which may be immobilized (as in IPG strips). The proteins are then denatured (using urea) and layered over the length of the gel (loaded). When a current is applied to either end of the gel, each protein will migrate to a unique

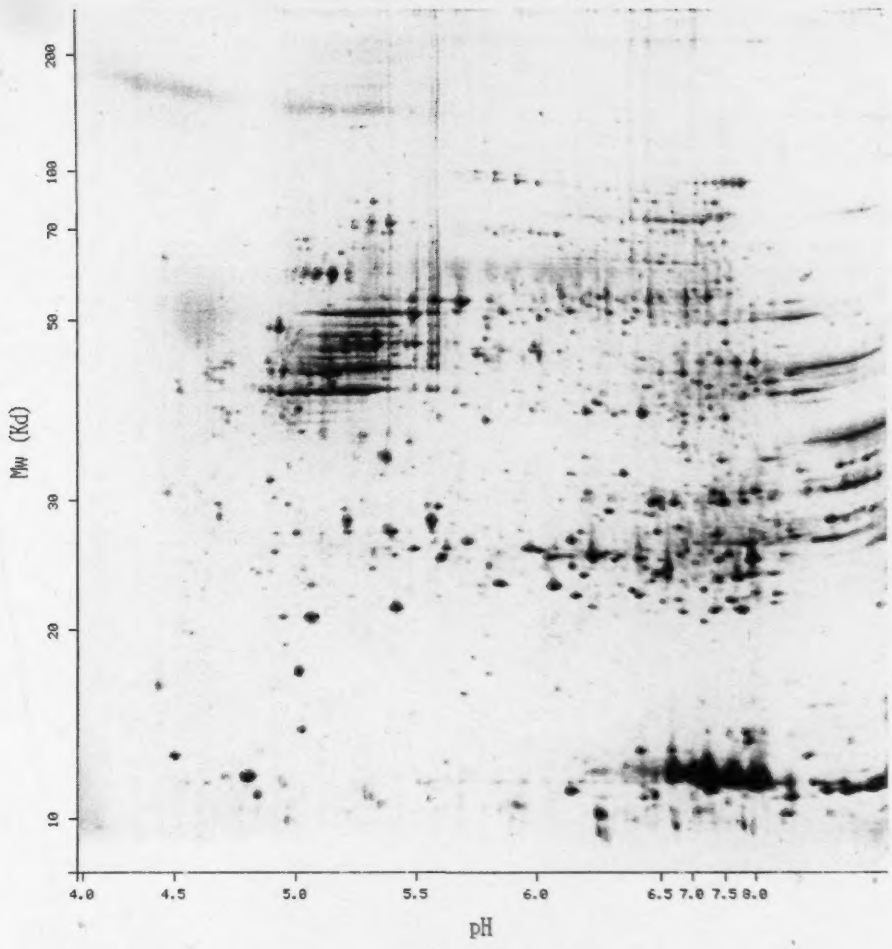


Figure 1.1 An Example of a High Resolution 2D Gel Acquired from SWISS-2DPAGE (http://us.expasy.org/cgi-bin/map2/noid?CEC_HUMAN). This human colorectal epithelium gel is performed by Department of Surgery and Pathology, University of Erlangen, Germany.

position corresponding to its unique pI. IEF is normally the first step or first dimension in 2-DE.

In the second dimension, which is perpendicular to the first one, proteins are typically separated by SDS-PAGE (Sodium Dodecyl Sulfate Polyacylamide Gel Electrophoresis) according to their molecular weight. In SDS-PAGE proteins are first treated with the detergent sodium dodecyl sulphate (SDS). The SDS molecules coat the protein with negative charges in proportion to the protein mass. The next step in SDS-PAGE is to put the proteins on a polyacylamide gel, and to apply an electric current. Because of their SDS coating, all the proteins will move through the gel towards the anode. In SDS-PAGE, the gel matrix is composed of a mesh of cross-linked acrylamide which provides hydrodynamic resistance in proportion to the level of gel crosslinking and the size of the molecule that is being moved through the gel. In other words, the bigger mass a protein has, the greater the resistance it will feel, and the slower it will move. Thus, by using SDS-PAGE in the second dimension of 2-DE, we can separate proteins according to their molecular weight.

To visualize protein spots on 2-DE gels, proteins are usually stained with organic dyes or metal ions. However, the number of suitable candidates for staining is limited by the nature of the separation methods and the requirement for subsequent MS analysis. For instance, since IEF separates proteins according to

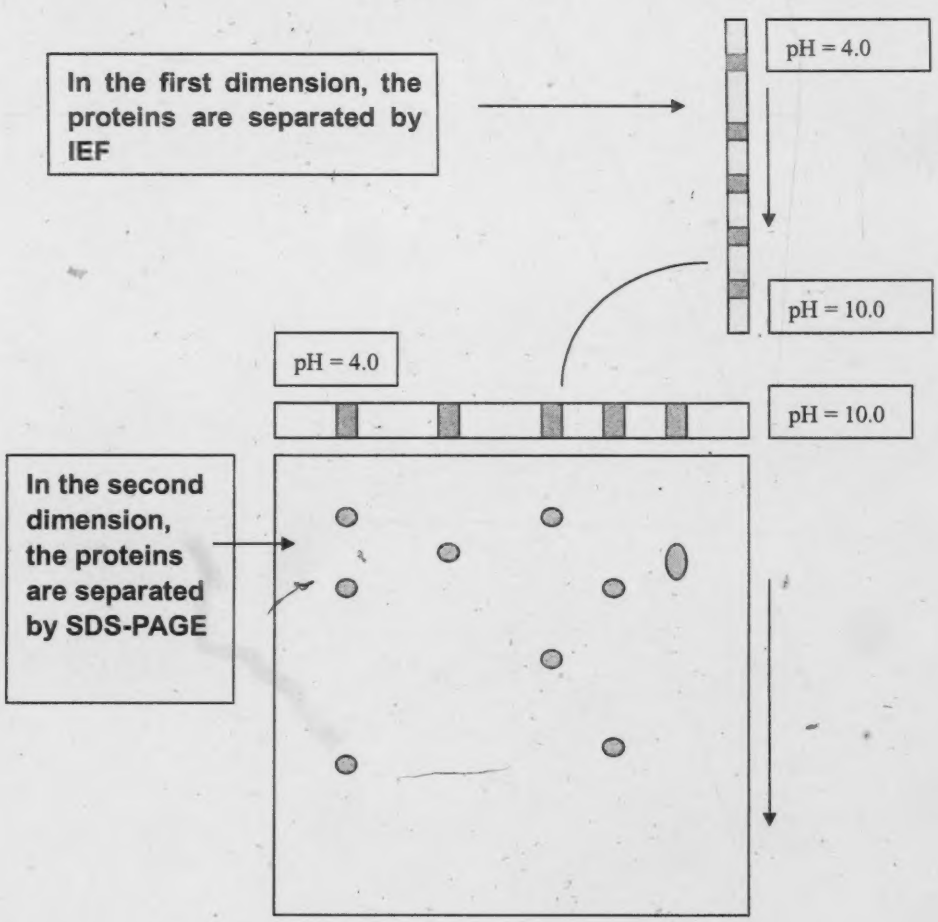

In the first dimension, the proteins are separated by IEF
pH = 4.0
pH = 10.0
pH = 4.0
pH = 10.0
In the second dimension, the proteins are separated by SDS-PAGE

Figure 1.2 A Schematic Illustration of The Two-dimensional Electrophoresis Process. Proteins Are Separated According to Charge and Molecular Weight.

pI, any pre-labeling method that changes the charge state of proteins, either by removing or adding a charged group, is not compatible with 2-DE. Likewise, since SDS-PAGE is conducted in a harsh chemical environment (SDS is a negatively charged detergent), these conditions can seriously compromise the performance of noncovalently bound dyes. Because of the limitations above, silver staining (Rabilloud et al., 1994) and Coomassie Blue staining (Schagger et al., 1988) are the two methods being most widely adopted in 2-DE gel laboratories. Both labeling methods are compatible with MS analysis. In silver staining, the silver ions can bind to amino acids side chains, such as the sulfhydryl and carboxyl groups of proteins (Coligan et al., 1995). Under suitable conditions, silver staining can allow one to visualize as little as one nanogram of protein. However, because of its multi-step complexity, the gel-to-gel reproducibility is somewhat wanting and variations of 20% in spot intensity having been documented (Quadroni and James, 1999). Relative to silver staining, Coomassie Blue staining is much simpler, quicker and costs little. However, its limit for detection is 8-10 nanograms (Patton, 2002) and the gel-to-gel reproducibility is also hard to control. Since SDS-PAGE separates proteins according to their molecular weights, proteins with known molecular weight can serve as standard molecular weight markers. For example, triose phosphate isomerase, horse heart myoglobin, bovine α-lactalbumin, bovine aprotinin, bovine insulin chain B, and bradykinin are frequently used as molecular weight standards (Schagger and von

Jagow, 1987). These molecular weight standards are often run in a separate "lane" in 2D gels to assist in MW calibration.

After the gel is run and stained, protein spots/bands from the gel can be cut out and digested by trypsin. The resulting digest can be subjected to mass spectrometric analysis, which will generate peptide-mass spectra profiles corresponding to the protein spots (Shevchenko et al., 1996). By searching specifically developed sequence/mass databases that contain peptide mass fingerprint information (Perkins et al., 1999), researchers can identify the proteins found in the selected spot or band. Peptide mass fingerprinting was adopted in the research described in the chapter 4 of this dissertation in which MS analysis was used to confirm the purification of the M.t. 0776 protein.

1.5 Bioinformatics and Drug Discovery

In the last few decades, advances in molecular biology have allowed the rapid sequencing of large portions of the genomes of several species. In fact, to date, more than 121 bacterial genomes, as well as 19 eukaryotic genomes (including yeast, *C. elegans*, drosophila, mouse and humans) have been fully sequenced. This enormous quantity of sequence information (Figure 1.3) along with other kinds of biological data (2D gels, X-ray structures, microarrays, etc.) has necessitated its careful storage, organization and indexing using computers and computer databases. This need for biological data management led to the

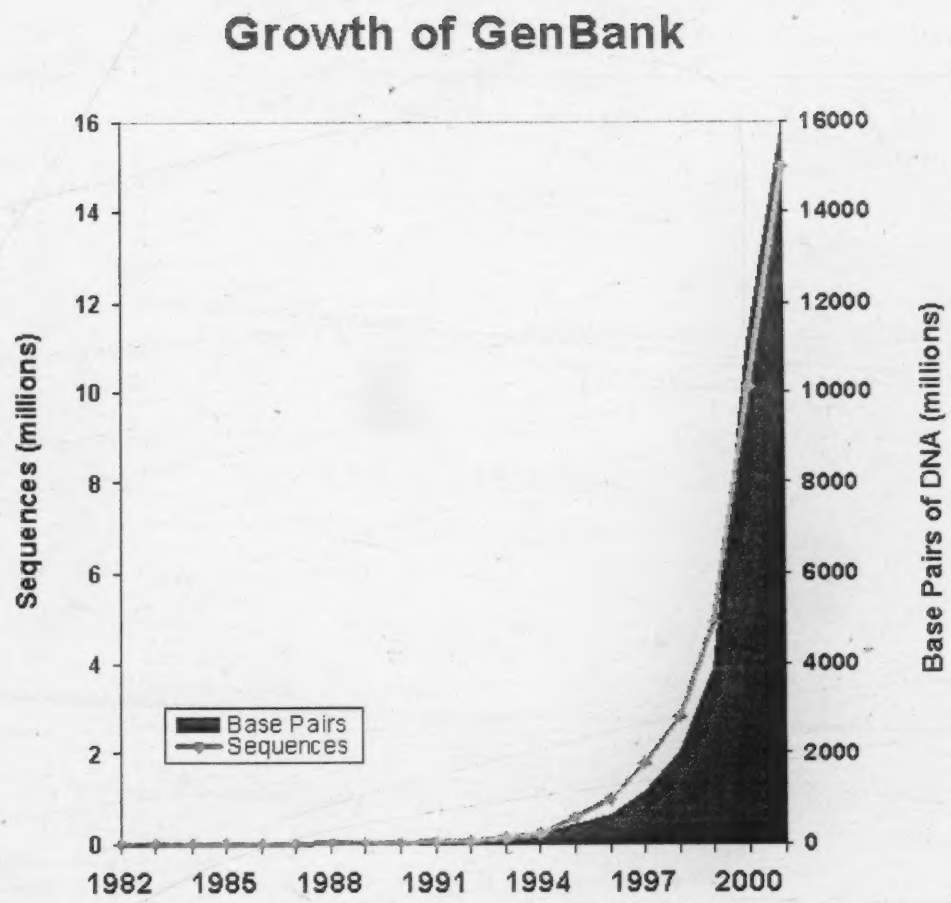


Figure 1.3 The Exponential Growth in the Number of the Total Sequences Deposited in GenBank (http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html)

development of a new and highly specialized area of life science called bioinformatics. Simply stated, bioinformatics is a field of information technology which endeavors to improve the storage, management and analysis of biological data. Practically speaking, it is concerned with the handling and analysis of DNA and protein sequence data. Bioinformatics is finding its way into many areas of life science research and is beginning to have an important impact in genetic disease research, evolutionary biology, cancer research, basic biology and even health economics. Bioinformatics is also among the areas that many in the pharmaceutical industry are betting will have a significant impact on the drug discovery and development process. In particular, bioinformatics is expected to contribute to the speed and efficiency of: 1) (in silico) drug target discovery; 2) drug target validation; 3) rational drug design; and 4) drug toxicology or drug safety assessment. In the areas of drug target discovery and validation, bioinformaticians are developing powerful sequence searching tools (such as BLAST and PSI-BLAST – Altschul et al., 1997) and sequence analysis tools (such as PSORT and TMHMM – Nakai and Horton, 1999; Krogh et al., 2001) which are allowing pharmaceutical researchers to identify the probable functions, active sites, cellular locations, substrates and structures of thousands of newly sequenced proteins in a matter of minutes. This process of computational target identification and validation is saving thousands of hours of painstaking experimental work.

With the knowledge the protein targets, bioinformatics can also help in rational drug design. A growing collection of bioinformatics tools for protein structure prediction (Homodeller and 3D-PSSM - Kelley et al., 1999; http://redpoll.pharmacy.ualberta.ca/- homodeller/) along with a growing database of solved structure (>20,000) is allowing pharmaceutical researchers to generate 3D structures of many protein targets in a mater of seconds or minutes. If a protein target's structure is known, researchers can artificially "dock" drug candidates to the targets using bioinformatics software and tools. An outline of this "new" drug discovery and design paradigm which is increasingly being adopted by the pharmaceutical industry is shown in Table 1.1.

Table 1.1. The New Paradigm in Drug Discovery and Design (Wishart, 2000)

Wet Chemistry	Computer chemistry
Gene Sequencing	Bioinformatics
Protein Expression/purification	Bioinformatics
Structural Analysis	Structure Predication
Lead Compound ID	Cheminformatics
Lead Compound Screening	Docking/Cheminformatics
Lead Compound Refinement	Docking

Among the many potential applications of bioinformatics in drug discovery, one of the most important is in data integration. The volume of genomic and proteomic information now being generated (~ 1 terabyte a day) has made it necessary for most proteomics researchers to use computers and special database

systems to handle this data. But data alone is not knowledge. To make any kind of data useful, it must be converted to knowledge by organizing and linking it to known facts or theories. This linkage and organization is best achieved through data integration.

An excellent example of data integration in proteomics is called ExPASy (**Ex**pert **P**rotein **A**nalysis **Sy**stem), which can be accessed at http://www.expasy.ch. It provides comprehensive databases, such as SwissProt which contains protein sequence data, detailed descriptions of protein function, domain structure, post-translational modifications, variants, and so on. Moreover, ExPASy provides links to other important biomedical databases such as Genbank (Benson et al., 2002) for gene sequence information, PDB (Berman et al., 2000) for structure information and GeneCards (Rebhan et al., 1997) for function and disease information. ExPASy also supports SWISS-2D PAGE, a comprehensive 2-DE gel image archive which links protein expression data with the many other databases in ExPASy. This level of data integration allows researchers to seamlessly connect expression data with disease data, sequence data with structure data or putative drug target data with disease information.

In chapter 2 and 3 of this dissertation, I will describe some of the bioinformatics tools I have developed to explore new ways to facilitate data integration, for both proteomics and for drug discovery. In particular in chapter 2, I will describe a

novel, web-enabled 2-DE gel archiving and viewing system that integrates protein expression data with protein sequence and physical property data. Additionally, in chapter 3, I will describe a new type of integrated drug database designed to facilitate drug discovery by linking protein target data to small molecule drug information. The motivation for the work in Chapter 3 is described in the following section.

1.6 Drug databases

Each year, the FDA approves 10-20 new drugs and 200-300 generic drugs for market release (Figure 1.4) (http://www.fda.com). A quick inspection of the Health Canada drug list (http://www.hc-sc.gc.ca/fnihb/nihb/pharmacy/-drug-benefitlist/) shows that in Canada alone, there are 22,000 approved drugs with 5200 of these being prescription drugs. The volume and variety of drug information is now rapidly approaching the volume and variety of data found in proteomics and genomics research. Just as bioinformatics has emerged as a field to handle biological data, drug informatics is now emerging as a field to handle drug data – especially for pharmaceutical scientists, pharmacists, doctors, and patients.

In recent years, a number of online drug databases have appeared, including RxList (http://www.rxlist.com), WebMD (http://www.webmd.com), and

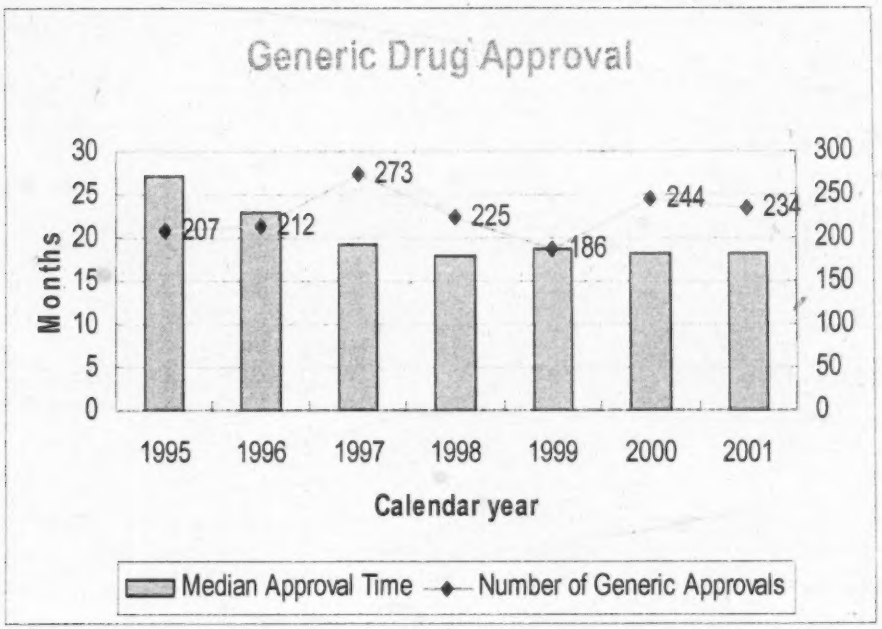


Figure 1.4 FDA generic Drug Approval Rates (months to approval and number of generic approvals) for the Past 7 Years.

MEDLINEplus (http://medlineplus.gov/). RxList is a comprehensive searchable database of information on approved or near-approved drugs. It is a free service to the internet community. Drug descriptions also include discussions of recent clinical trials. The database is searchable and provides a description of each drug along with its generic and trade name, clinical pharmacology, indications, contraindications, dosage and administration, adverse reactions, and storage recommendations. Monographs on each drug are derived from Mosby's GenRx. Rxlist contains the generic name, brand name, and therapeutic category over 4,000 U.S. prescription and over-the-counter (OTC) medications. Several categories are now hyperlinked to listings of indications and side effects. MEDLINEplus is part of the National Library of Medicine (the world's largest medical library). It is a guide to prescription and OTC drugs provided by the United States Pharmacopeia. One of the links on MEDLINEplus is "Drug Information", which allows users to browse by brand or generic drug name for information on more than 9,000 prescription and over-the-counter drugs. The page also links to the latest drug news from the U.S. Food and Drug Administration. This site also contains drug brand names, category, side effects, usage and dose.

As we can see from the above descriptions, most online drug databases provide a good deal of clinical information about drugs, such as prescription information, indications, side effects, and pharmacology. In essence, these drug databases

mainly serve patients and doctors. However, there are as yet no drug databases that link pharmacological or clinical data to the vast wealth of chemical, molecular, genomic or biological data about either the drug or its protein drug target. This is an area of research – linking of drug informatics to bioinformatics and proteomics – which I have attempted to address in Chapter 3 of this thesis.

In this chapter, I describe a freely accessible, web-enabled, queryable database (DrugBank) that links drug structure and activity data with protein structure, function and sequence data. DrugBank brings well-developed bioinformatics concepts of search and comparison to medicinal chemistry-linking bioinformatics, proteomics and drug discovery together. As such, it can serve as an online teaching tool for medicinal chemistry, an information hub for patients or professionals, a drug hotline or a drug discovery "system".

1.7 Structural Proteomics

Simply stated, structural proteomics aims to generate useful three-dimensional (3-D) structures of entire proteomes by a combination of experimental structure determination and modelling. The most straightforward way to predict the 3D conformation of a protein is to use a closely related (>30% sequence identity) solved structure as a template, and align the query sequence to the sequence of a template structure. This is called "comparative" or homology modeling (Figure 1.5). By detecting structural homology with proteins of known function,

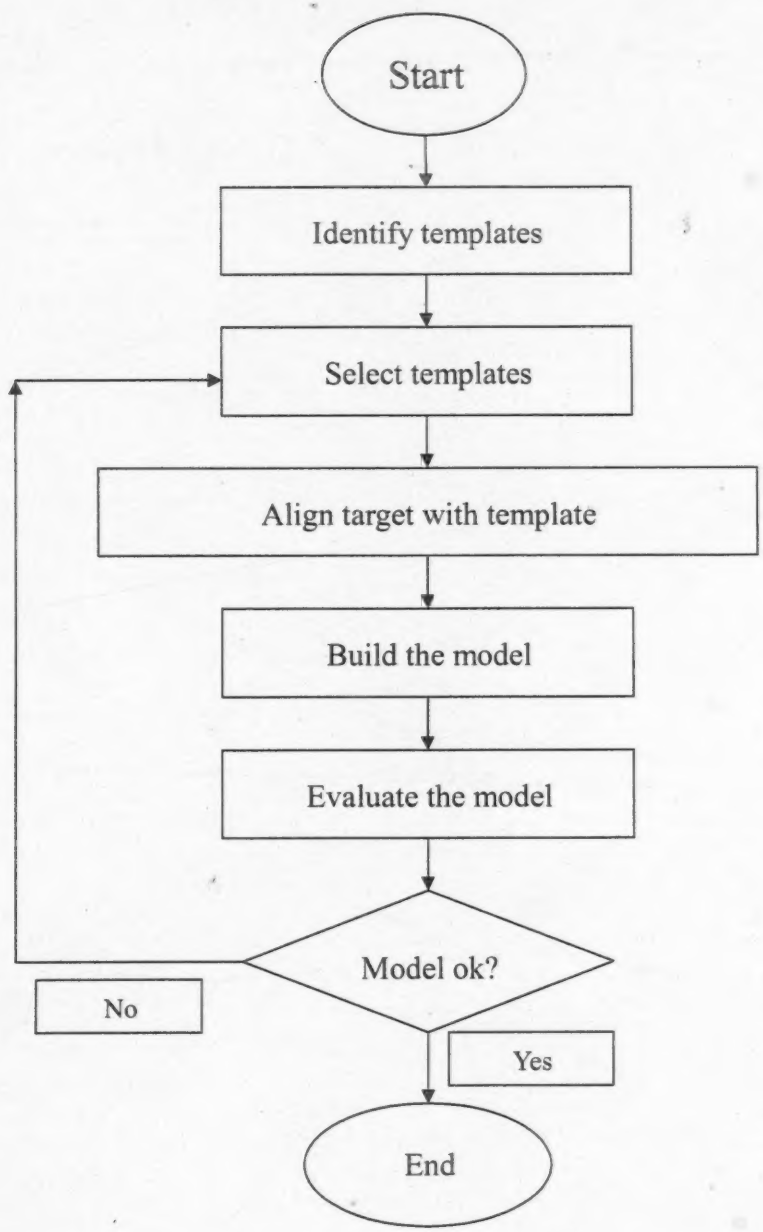


Figure 1.5 Procedures for Comparative Modeling

researchers may identify or assign protein functions to new proteins that were unachievable from sequence analysis (Zarembinski et al., 1999). Another role for structural proteomics is in helping to solve the protein folding problem. The idea is to solve a sufficiently large number of different 3D protein structures to serve as a "basic parts list" of protein folds (Sali, 1998). By using the parts list and a variety of computational (comparative modeling, threading, etc.) technologies, other protein structures can be modeled. That is, structural proteomics could be used to solve the protein folding problem by "brute force".

However, the limited number of available 3-D protein structure templates currently restricts scientists from making large-scale functional annotation using structural information. Recently, it has been estimated that the number of structures needed to model 90% of the sequences, at >30% sequence identity is approximately 16,000 (Sundström, 2001). Based on available genome data and on the inventory of unique structures from the Protein Data Bank (PDB), it was also shown that structural templates were available for segments of only 30% to 35% of all known genome sequences. Clearly, we have a long way to go before the dream of knowing every protein structure is realized.

There are currently several promising structural proteomics projects around the world. In the United States, nine pilot centres to develop high-throughput pipelines for structure determination were funded by the Protein Structure

Initiative of the National Institute of General Medical Sciences (NIGMS). These projects include Berkeley Structural Genomics Center, Center for Eukaryotic Structural Genomics, The Joint Center for Structural Genomics, The Midwest Center for Structural Genomics, New York Structural Genomics Research Consortium (NYSGRC), Northeast Structural Genomics Consortium (NSGC), the Southeast Collaboratory for Structural Genomics (SECSG), Structural Genomics of Pathogenic Protozoa Consortium (SGPPC) and the Tuberculosis Structural Genomics Consortium. A protein target database, TargetDB (http://targetdb.pdb.org), has been constructed by the Protein Structure Initiative group based out of NIGMS. This database contains the records of protein being expressed, crystallized or solved by all consortium members. More than 5000 difference proteins are currently in the database. In Europe, there are a number of promising structural proteomics projects including the Protein Structure Factory (PSF) initiative coordinated at the Max Delbrück Center for Molecular Medicine in Berlin, Germany, as well as the NMR software initiative at the University of Utrecht (Heinemann, 2000). In Japan, structural proteomics research is thriving too. The largest projects are the RIKEN Structural Genomics Initiative, and the MITI-funded structural genomics project at the BIRC (Yokoyama et al., 2000).

The Ontario Cancer Institute (OCI) is currently leading another structural proteomics project to determine the feasibility of large-scale structural biology (Yee et al., 2003). There were three primary purposes of this study: 1) to evaluate

the difficulties researchers could meet in such a high throughput project; 2) to estimate the percentage of proteins encoded by a genome that are immediately amenable to structure analysis; 3) and to assess the extent to which function can be inferred from structure (Christendat et al., 2000). This project started in 1998 and has led to the cloning and expression of over 700 proteins of unknown structures and functions from *Methanobacterium thermoautotrophicum* (M.t.) *ΔH*. Membrane proteins and proteins with significant sequence similarity (BLAST search with an e-value cutoff of 10^{-4}) to proteins in Protein Data Bank (PDB) were specifically excluded in the project. The cloned proteins have been "farmed" out to dozens of X-ray crystallographers and NMR spectroscopists across Canada and the US.

M.t. 0776 is one of the proteins selected by the OCI team for structural proteomics analysis. This protein has no known function or similarity to any other protein sequences. In chapter 4 of this dissertation, the expression, purification, and preliminary characterization of this small protein will be described.

1.8 Outline of this Dissertation

This thesis consists of 5 chapters. The first chapter provides the background and context to the work and results described in the next 4 chapters. Chapter 2 describes the web-based gel analysis tool called GelScape while chapter 3 describes the design and implementation of the drug-protein database called

DrugBank. Chapter 4 describes the expression, purification and preliminary characterization of a protein (M. t. 0776) of previously unknown function while Chapter 5 provides a general discussion and conclusion.

1.9 My Contribution to this Dissertation

The work described in chapter 2 and 4 of this thesis, including the design, data collection, testing, isolation and wet-bench work was done almost entirely by me, with occasional advice and periodic assistance provided by my supervisor and colleagues. With regard to the work described in chapter 3 (DrugBank) however, it is important to point out that this is a very large and collaborative project involving the contributions of several part time staff and summer students over a period of ~2 years. Because so much of the data in DrugBank is not electronically accessible or linked, the data entry component represents one of the most significant and difficult tasks in this project. For the data entry part, as well as the data validation and formatting, I would estimate I contributed ~40% to this component. Jennifer Woolsey and Mahin Habibi-Nazhad contributed to the rest of data entry. In order to make the database accessible and queryable for users, web pages and CGI programs were needed to make it functional. Once again, a number of colleagues in Dr. Wishart's lab helped me implement some of these components. The web page set-up, drug card formatting, CGI programs for retrieving "DrugCards", index files for text queries as well as CGI scripts for the structure query tool and data extractor were written entirely by me. Dr. Anchi Guo

assisted me in implementing the DrugBank data extractor. Dr. Bahram Habibi-Nazhad and Dr. Anchi Guo contributed ~70% of the DrugBank structural query code, and I was responsible for the rest. Haiyan Zhang assisted me in implementing the BLAST search function on DrugBank. In addition to the data entry and CGI programming, I have been responsible for the most of the testing, data checking and error correction in DrugBank.

With regard to the work in chapter 4, I was responsible for developing and refining the transformation, preliminary expression and purification protocols. All the following bioinformatics analysis and predictions were conducted by me. I was assisted in the acquisition of 1D NMR data and MS data by several members in Dr. Wishart's laboratory.

Chapter 2: GelScape

2.1 Introduction

Despite the predictions of its imminent demise, 1D and 2D gel electrophoresis (GE) is and will likely continue to be an essential cornerstone to proteomics research. No other low-cost technology has comparable resolution or sensitivity for protein separation and display. The ability to digitally scan and archive 1D and 2D gels has given new life to gel electrophoresis as now computers can be used to manipulate, annotate, store and compare gels in ways that were impossible even a decade ago. Today, there are a number of popular, commercial software packages that are available to facilitate digital gel analysis, including Melanie 4 (GeneBio), Phoretix 2D (Phoretix Inc.), ImageMaster 2D (Amersham), PDQuest (BioRad), Z3 (Compugen), ProXPRESS (Perkin-Elmer) and Gellab II (Scanalytics). These stand-alone programs support an impressive array of interactive image analysis and annotation routines including automated or semi-automated spot annotation, spot integration, gel warping, image resizing, HTML image mapping, image overlaying as well as the storage of gel image and gel annotation data in compliance with Federated Gel Database (Appel et al., 1996) requirements. However, the high cost and restricted platform compatibility of many of these commercial packages have led several groups to develop web-based freeware or groupware systems to handle certain key aspects of gel analysis, including gel comparison (Flicker – Lemkin, 1997) or interactive

gel exploration (WebGel – Lemkin et al., 1999). Likewise, a growing number of interactive gel archiving or gel viewing websites (databases) have appeared including: SIENA-2DPAGE (http://www.bio-mol.unisi.it/2d/2d.html), SWISS-2DPAGE (http://ca.expasy.org/ch2d/) and a 2-DE database by Algorithmus, Prehm Berlin (http://www.mpiib-berlin.mpg.de/2D-PAGE/). A more complete list of gel database web sites can be found at http://ca.expasy.org/ch2d/2d-index.html. With the development of Java and the appearance of Java applets, Java servlets, Dynamic HTML (DHTML), XML and other web languages, a new generation of internet tools is beginning to provide methods for the interactive manipulation of image data over the web. Given the growing trend towards web-based bioinformatics systems and the growing importance of 1D and 2D gels in proteomics, we decided to try and exploit these new internet tools and develop a web-based software system that could perform at least some of the gel or image analysis functions found in commercial software packages.

As a first step towards this goal, we decided to investigate the application of Java-based image manipulation tools towards developing more useful 1D and 2D gel databases. While many of the gel databases already mentioned offer simple HTML image maps (i.e. clickable images) for interactively viewing gel spots and bands, none of them provide sufficiently advanced tools for users to fully interact with the gel images. Given that many 2D gels can contain over 3000 protein

spots, there is a clear need for proteomics researchers to be able to access tools that can help them selectively zoom in and out the gel image, to rapidly identify gel spots or, with the touch of a mouse button, provide detailed information for a selected spot. To meet this need, a software tool called GelScape was constructed. GelScape is a web-based gel archiving system which enables users to interact with archived 2-DE gel images *via* Java applets, Javascript modules, HTML, Perl and Dynamic HTML. Specifically GelScape offers five unique features or tools: 1) a hyperlinked navigation window; 2) an image-mapped gel viewer; 3) a rubber-band zooming tool; 4) a gel magnifying glass tool; and 5) a gel comparison tool. GelScape also has a collection of pre-annotated gel images. A more complete description of GelScape and the GelScape database follows.

2.2 System and Methods

2.2.1 Platform and Availability

Web pages are normally written in HTML (hypertext markup language). HTML provides the graphical front-end to a web page. However, to make a web page "come to life" so that it can respond to or accept user input, one must write a special back-end program using the conventions developed for the Common Gateway Interface (CGI). The CGI scripts for uploading and handling gel images were written in PERL v5.6.1 (Wall et al., 1996) on a SGI computer running UNIX. The web interfaces were written in HTML, Dynamic HTML and Javascript on a desktop computer running Windows 2000®. The complete

package was ported to a LINUX server and is now running on Dr. Wishart's web/file server called redpoll. GelScape can be accessed freely at http://redpoll.pharmacy.ualberta.ca/~zchang/GelScape/ -index.html. GelScape can be viewed on any computer with a Netscape browser (version 3.0 or later) or running Internet Explorer (version 4.1 and later).

2.2.2 Data Sources, Storage, Retrieval and Display

Currently, there are ten pre-annotated gels in the GelScape database. They include a 2D gel of human heart atrium proteins -- acquired by Berlin HEART-2DPAGE (Virchow Clinic of the Humboldt University Berlin, German Heart Institute, Berlin), a 2D gel of human breast epithelial proteins provided by Argonne/PMG (Center for Mechanistic Biology and Biotechnology at Argonne National Laboratory), a 2D gel of human breast MCF7 variants proteins by LMP/NCI (U.S. National Cancer Institute), as well as four different 2D gels of human (cerebrospinal fluid) CSF, human liver, human plasma and human kidney – all provided by ExPASy. These 2D gels constitute a small but useful protein database of proteins found in human body fluids and tissues. The gel images and gel annotation information were stored using standard HTML conventions. Each spot (which has been previously identified by MS analyses) on a given gel page is hyperlinked to a "protein card" which appears when the user clicks on a gel spot. Spot positions for each gel in the GelScape database were mapped to the gel image using standard image mapping software (see later). The protein card is a

synoptic summary providing detailed information on the protein's name, sequence, isoelectric point (pI), molecular weight, and a link to the SWISS-PROT sequence database (Figure 2.1).

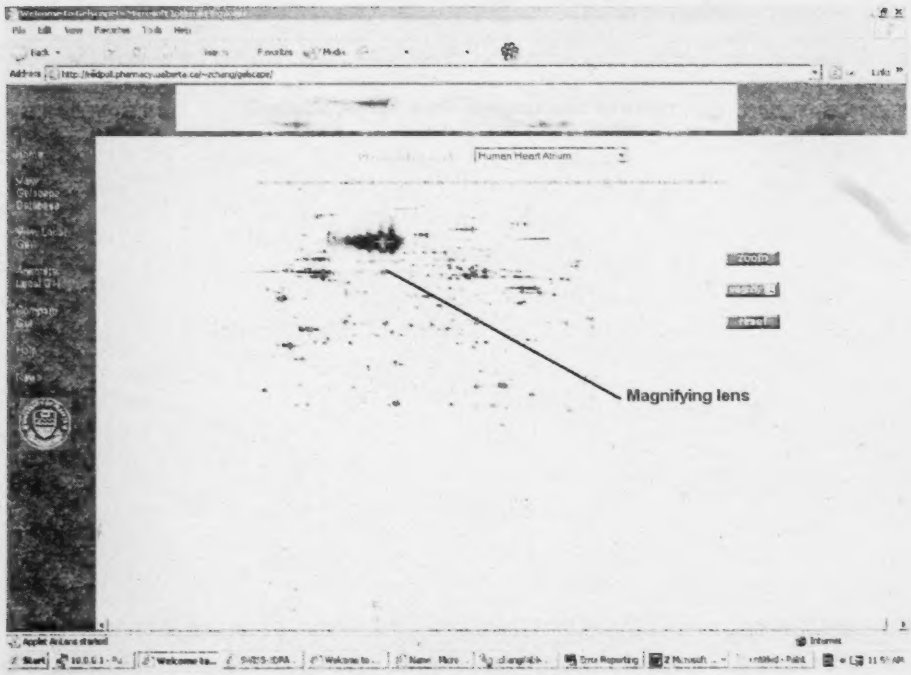
To facilitate gel image viewing and user-interaction, GelScape displays its information using three different HTML frames (Figure 2.1). Frames are an HTML construct that allow users to independently browse or view different images in different windows, while at the same time allowing seamless communication between other frames or windows. The left-margin or border frame contains the GelScape menu or navigation tools that allows users to select various GelScape options. The upper border frame contains the GelScape title (images and text) and the largest or central Gel View frame displays the gel images as well as the results of any interactive operation.

GelScape was designed to enable the interaction between users and gel images. Java™ applets were used to make the interaction possible. Two Java applets were implemented, a magnifying glass (lens) applet and a so-called "rubber-band" zooming applet. The magnifying lens applet magnifies images over a pre-defined region (50 pixels in diameter on the original image) and overlays that magnified image directly over the primary image when a user mouses over the gel image. This applet was modified from Fabio Ciucci's AnLens (http://www.anfyteam.com/). The rubber-band zooming applet allows users to

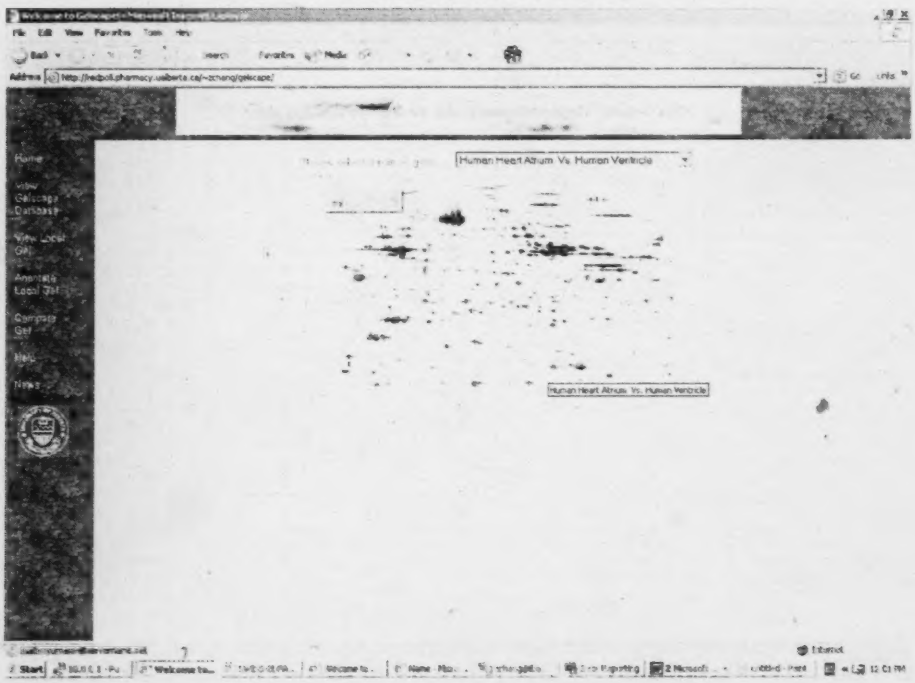
A. GelScape Homepage

B. GelScape Protein Card

Figure 2.1(a) The GelScape Layout.



C. Applet tool for viewing gel images



D. Gel comparison on GelScape

Figure 2.1 (b) The GelScape Layout

draw out a rectangular region on the original gel image and to have that region expanded and viewed over the original gel image. The zooming applet was modified from one coded by Vivaorange (http://www.vivaorange.com/). Because the applets can enlarge gel images, the default image size on all GelScape pages was set to 450 x 300 pixels. This image size is sufficiently small to accommodate computers with minimal monitor resolutions of 600 x 800 pixels.

2.3 Implementation

GelScape consists of multiple HTML pages containing HTML, DHTML, Java scripts and Perl CGI scripts. It also contains gel images (GIF format), image map coordiantes and gel data about individual spots on each gel. Each of these components is linked or connected through a structured menu maintained on a centrally maintained webserver. An overview of GelScape's implementation is shown in Figure 2.2.

2.3.1 General GelScape Interface and Home Page

As already mentioned, GelScape's navigation frame was placed on the left side of each GelScape page including the GelScape homepage. Using this navigation menu, users can interact with the GelScape server and choose the particular page or GelScape resource they wish to go to. This menu or navigation tool provides six menu options including: 1) View GelScape Database; 2) View Local Gel; 3) Annotate Local Gel; 4) Compare Gel; 5) Help; and 6) News. The GelScape

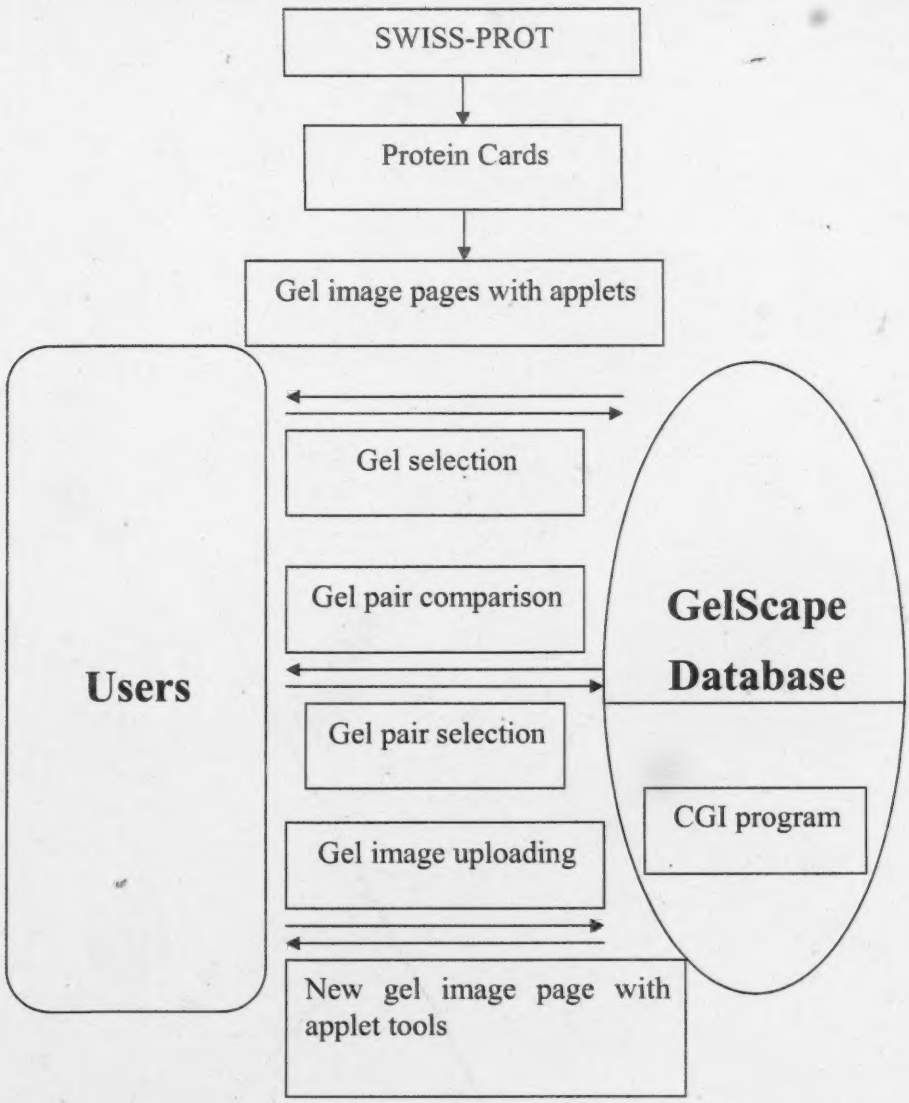


Figure 2.2 Overview of the GelScape Implementation

database is intended to be the primary destination for most users. By default, only one gel image is displayed after the user chooses to view a gel within the GelScape database. At the top of the page is a pull-down menu which indicates the name or source of the gel being viewed. Other gels can be selected for viewing through the pull-down menu. Each gel is marked with red "+" signs indicating the location of image-mapped protein spots. Image mapping in HTML was used to generate the clickable spots on each gel and Java scripts were used to make the pop-up protein cards. On the right side of each gel image are three buttons: 1) Zoom; 2) Magnify; and 3) Reset. Each gel image is zoomable when users click on the "zoom" or "magnify" buttons, which are always located in the central Gel View frame. These image manipulation options were realized by a few line of codes in HTML and Java script. To undo a zoomed image or to turn of the magnify applet, users must press the Reset button.

2.3.2 GelScape Image Maps

For each identified or marked protein spot, an HTML document (a protein card) containing that protein's information (pI, molecular weight, and protein sequence) was created using Microsoft FrontPage® 2000. The protein card information was collected from recognized sources such as SWISS-PROT (Boeckmann et al., 2003), the Protein Data Bank (Berman et al., 2000) and GenBank (Benson et al., 2002). To create an image map for each gel, a clickable area corresponding to each identified spot was created on the gel image which was also marked with a

red "+" sign. Image mapping and spot marking were implemented so that users could easily find the spots of interest and obtain relevant protein information with the click of a mouse button. This feature was realized by using the standard image map functionality of HTML (as implemented in FrontPage). The area around the known protein spots can be made clickable by using the coordinates of the spots on the gel image. When the "hot spots" are clicked, the hyperlinks related to the spots lead users to the HTML documents (protein card) containing the information about the selected spot.

2.3.3 Uploading and Comparing Gel Images

In order for the users to upload gel images onto GelScape's server and to view gels using GelScape's viewing tools, a common gateway interface (CGI) program was developed. This CGI program was required have the ability to identify the proper file format (ending with .jpg, .gif or .tif suffix) and the ability to generate the necessary web pages containing the gel image and the applet gel-viewing tools. It was also required to have the ability to prevent malicious abuse of the GelScape server.

In meeting the above requirements, a CGI program was written which performs most of these functions. The CGI "Upload" script first checks each file name to prevent users from accidentally uploading malicious code (computer viruses, Trojans, etc.) that may hide in the file name. If the file name passes this test, the

program checks for the existence and size of the file. If the file size exceeds 1000 kilobytes an error message is returned and the file will not be uploaded. If the file size requirement is met, the file will be uploaded and the gel-viewing HTML pages along with the appropriate applet tools will be generated. All uploaded files are saved in a "/temp/" directory on the redpoll (Dr. Wishart's laboratory's central web and file) server. Permissions are such that users cannot access other directories on this server. In this way, GelScape provides a safe environment for uploading local gel images to its server location.

To implement the gel comparison function in GelScape a "Mouse-over" Java script was constructed. To compare gels, two gels must first be selected and uploaded using the pull down selector box located at the top of the Gel View frame. By uploading both images to the same location, the two gel images are effectively placed on top of each other in the Gel View window. By moving the mouse repeatedly over and then off the visible gel image, the second gel can be made to appear or disappear. In this way the gels can be "flickered" at a rate which allows users to compare one image to another in a facile, user-controlled manner.

2.3.4 Applications

The strength of GelScape lies in the way it facilitates the interaction between users and gel images. Specifically, GelScape makes it easier to observe and

compare 2D gels. The zooming and magnification options allow users to zero-in on gel spots of interest, while the protein information found on each "protein card" provides users with most of the detailed information they need about a given protein spot. Likewise the gel comparison function allows users to readily identify differences and similarities between gels. For example, if one wished to study heart disease, one could look at the differing protein expression profiles between two patients (one with disease, one without) or between two different tissues (one affected, one not). For instance, differences between the proteomes of heart ventricle and heart atrium tissues may give researchers important hints about the proteins that play key roles in cardiovascular disease or myopathy. As can be seen in Figure 2.3, where we are looking at the annotated 2D gels of human heart atrium (A) and human heart ventricle (B) -- both gels are quite similar and both gels have numerous (>500) protein spots. Using GelScape, researchers can use its gel viewing function and select one or both gel images (human heart atrium or human ventricle). Using the magnification and zooming functions, users can find the differences and similarities quite easily. By clicking on a protein spot of interest, users can also find specific disease-related protein information. Furthermore, users can find the differences (or similarities) between the two gels by using the Gel Comparison function and flickering the two overlaid gels on and off.

A. Annotated 2D gel of human heart atrium

B. Annotated 2D gel of human heart ventricle

Figure 2.3 Annotated 2D Gels of Human Heart Atrium (A) and Human Heart Ventricle (B)

2.4 Discussion

2.4.1 Database Construction and Information Display

GelScape was constructed with due consideration for the five rules originally set out for constructing Federated 2D PAGE databases (Appel et al., 1996). These rules are listed below: 1) Individual entries in the database must be accessible by remote keyword search. 2) The database must be linked to other databases through active cross-references. 3) There must be a main index which allows all databases to be accessed through a single entry point. 4) Individual protein entries must be accessible through clickable images. 5) 2-DE analysis software must be able to access directly individual entries in any federated database.

In compliance with the rule one for a federated 2-DE database, GelScape can be browsed interactively *via* the web and the gel images can be selected or searched *via* the pull-down box located at the top of the Gel View frame. According to rule two, images and spots in the GelScape database are linked to the SWISS-PROT databank and accession codes for the Protein Data Bank, GenBank and PIR (Protein Information Resource) are included. In accordance with rule three of the Federated Gel Database guidelines, a bidirectional cross-reference can be established in the individual information of the tables in the HTML documents for each protein spot. In conformity with rule four, individual protein entries are accessible by clicking on a gel image. The protein names are displayed when users move the mouse over marked (a red +) protein spots. GelScape complies

with rule five by providing powerful zooming and magnification applets which can be accessed simply by clicking one of the Zoom or Magnify buttons in the Gel View frame.

2.4.2 Database Administration and Maintenance

One key advantage of GelScape's database system is that it retrieves much of its data from the internet, so that researchers who use this system need not worry about limited disk capacity to store each gel card for each spot. GelScape uses pointers and links to existing web databases rather than compiling the data centrally in a single local repository. Another advantage to GelScape is that it makes the sharing of gel information among remote or distant laboratories possible. This is because GelScape, as a web tool, exploits the easy, transparent connectivity of the web and the near-universality of web browsers and web access. Although other web-based 2-D gel databases are available, they all have some weakness. Basically, to maintain those databases, sophisticated skills in computer programming as well as processor-intensive operating systems with expensive software and hardware are required. In order to make GelScape easy to administer and maintain, we constructed an HTML document (a protein card) for each known protein spot, which minimized the cost of maintenance. Sophisticated programming was not required in GelScape's construction and maintenance, but the data-entry work was laborious. Compared to other 2-DE gel databases, such as SWISS-2DPAGE (Hoogland et al., 2000), the NCI 2DWG Image Meta-Database

(Lemkin, 1997) and the 2-DE database at the Max Planck Institute for Infection Biology (Pleissner et al., 2002), GelScape appears to provide more powerful and convenient gel comparison and zooming tools, but it is quite clear that the size and comprehensiveness of the GelScape database is significantly smaller than these large institutional repositories.

2.4.3 Future Directions

We believe that GelScape and the general concepts developed in implementing GelScape may eventually be applied to real proteomics research. A key limitation with the current implementation of GelScape is the inability to interactively annotate gels from within the GelScape environment. Therefore, in order to help users annotate their own gels, we intend to provide several types of annotation functions in the second version of GelScape (GelScape II) which is now being developed by Nelson Young in Dr. Wishart's laboratory. Similarly, to facilitate gel comparisons, it will be important to have ways of matching gel images through visual "registration" or image warping. This feature has also been added to GelScape II. The development of a spot search or protein name utility through the implementation of a relational database system would make the archival and retrieval functions in GelScape significantly more useful. Once again this is being added to GelScape II. Finally, it would be desirable to increase the size of the GelScape database and to integrate more protein data into the system. A CGI program could be developed to regularly update and collect

data from the internet so that the GelScape database (and protein cards) could be updated automatically.

2.5 Conclusions

2D gels continue to be important separation and display tools which can be used to measure gene/protein expression, protein activity and protein interactions. The knowledge provided by 2-DE may help us understand how biological systems are constructed and to learn more about the mechanisms that control them. Given the importance of 2D gels in proteomics and the need to make this data as public and "transparent" as possible, we have developed a web-based gel viewing and archiving system. With the addition of gel annotation and image manipulation utilities (to appear in GelScape II) we are hopeful that GelScape (or its successors) may serve as a prototype for a new kind of tool that will help researchers understand, share, view and archive 2-DE gels on any computer, in any lab, around the world.

Chapter 3; DrugBank

3.1 Introduction

Proteomics is a high-throughput science that promises to bring thousands of new protein drug targets to our attention. However, the identification of which proteins are promising candidates is still a painstakingly slow process. Furthermore, once a promising target protein or even a target organism is identified, there is little in the way of high throughput techniques to identify promising drug leads. Bioinformatics potentially offers one route to intelligently sift through the mountains of new protein data and to rapidly identify both drug targets and drug leads. In this chapter I will describe the development and prototyping of a new kind of integrated, web-enabled, bioinformatics database -- called DrugBank -- that attempts to link clinically or pharmaceutically relevant drug information with basic genomic or proteomic information. The idea behind DrugBank is to move bioinformatics databases and applications from the realm of basic biology into the realm of therapy and clinical medicine. DrugBank is designed to contain and link information about small molecule drugs (names, structures, formulas, physical constants, synthetic methods, spectra) to their large-molecule targets or receptors (protein sequences, gene sequences, gene locations, protein structures, protein functions, polymorphisms). These data are further supplemented with pharmaceutically relevant data on drug metabolism, metabolic by-products, pharmacokinetics, pharmacology, indications, dosage,

bioavailability, distribution and alternate drug names. By bringing together data from a wide variety of sources, many of which are not electronically available, I am hoping to create a database system that is more than just an electronic pharmacopeia, but a true research tool.

To make DrugBank as useful as possible to the widest audience possible, I (along with my colleagues in Dr. Wishart's laboratory) have developed and incorporated a number of web-enabled search, query and display tools. DrugBank supports boolean text queries, protein and DNA sequence searches, whole genome and proteome searches, small molecule structure queries, small molecule sub-structure queries, and protein-ligand docking queries. Users are also able to interactively edit, draw, display and submit graphical data (chemical structures, protein structures, chromosome locations, SNP-data, protein-protein interactions) using a variety of Java applets. In addition to these text and graphical query systems, DrugBank also offers the capacity to conduct intuitive relational queries against the database using a specially designed data "extractor" originally developed for the CyberCell database.

By combining a broad selection of proteomic, genomic, metabolic and pharmaceutical data with powerful analytical, predictive and query tools, DrugBank could benefit medicinal chemists, molecular biologists, physicians, pharmacists, students and patients. It is hoped that DrugBank would allow a

wide range of "virtual" experiments to be done over the web. For instance, using DrugBank new chemical structures could be designed and "virtually" tested over the web, medicinal chemists could use DrugBank to quickly scan for structural similarities for newly synthesized drug leads, new genome sequences could be quickly scanned for the existence of proteins having homology to known drug targets, new insights could be gained about the action or physical properties of certain classes of drugs, or important data on deleterious polymorphisms for certain drug classes could be rapidly extracted and passed on to patients or physicians.

In the following pages I will describe in detail the data, the programs, web constructs and visual interface designs that have gone into constructing the first draft of DrugBank. In its current form, DrugBank contains 256 common prescription drugs from 131 drug categories. With the inclusion of brand names and various mixtures DrugBank actually provides information on over 3800 drugs.

3.2 System and Methods

3.2.1 Platform, Software and Availability

The programs used to construct DrugBank and web-enable the database consist of a large number of Common Gateway Interface (CGI) scripts written in PERL v5.6.1 (Wall et al., 1996). All programs and associated DrugBank data were

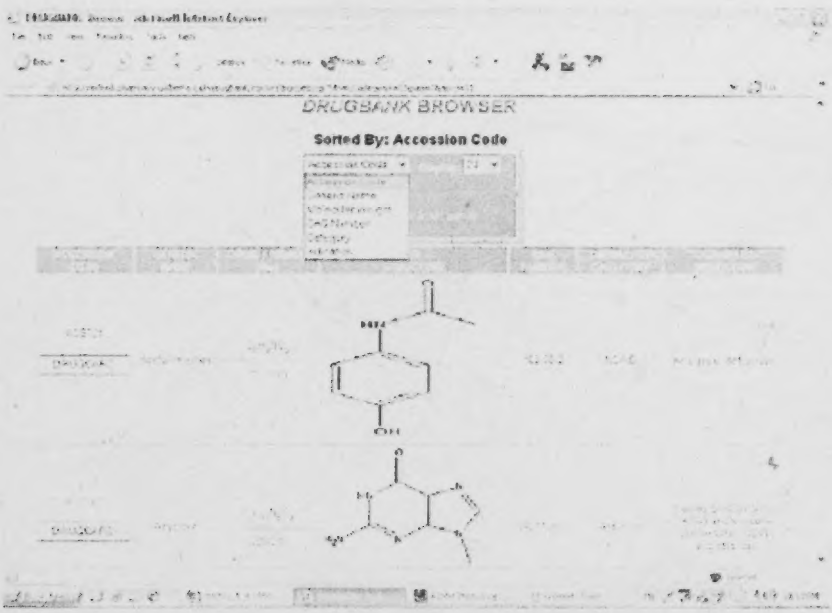
installed on an Apache 1.27 server on a SGI computer running Redhat Linux 7.3. Data entry was conducted on a personal computer running Microsoft Windows® 2000. Chemical structures of drugs were drawn using ChemDraw Ultra® 6.0 and saved in both JPEG (Joint Photographic Experts Group) and MOL formats. 3D structures were generated using Chem3D Ultra® 6.0 and saved in PDB (Protein Data Bank) format. SMILES strings were generated using BABEL (Walters, P. and Stahl, M., 1992) which converted MOL files directly into SMILES strings. DrugBank can be freely accessed at http://www.drugbank.ca/. A screenshot of the DrugBank homepage is shown in Figure 3.1.

3.2.2 Data Sources and Storage

Essentially all of the data in DrugBank was entered and stored in a simple flat file (tab delimited text) format including text files, MOL files, Protein Data Bank file (PDB files) and image files. The flat file format is a popular file format for biological databases because they can be easily parsed and formatted by simple programming languages such as PERL and additional (expensive) software packages (i.e. relational database systems such as Oracle) were not needed to maintain the database.

To create a medically and pharmaceutically useful drug database, quantity, precision and accuracy of information are all critically important. In order to ensure that the information in DrugBank was typographically correct, all data was

A. DrugBank Homepage



B. The DrugBank Drug Browser Window

Figure 3.1 The DrugBank Homepage and the Drug List

entered and then checked or re-entered by several different individuals, two of whom had BSc or PharmD training. Textual data and facts were collected from a variety of internationally recognized reference texts or chemical information sources. The initial set of 200 drugs in DrugBank was selected from a list of the most commonly prescribed drugs for 1996 (Pill Book, 9th Ed.) as supplied by the FDA (Food and Drug Administration). The indications, drug names, brand names, brand name mixtures, and pharmacological data were obtained from three main sources: the Merck Index (12th edition; Budavari, editor emeritus; editorial staff O'Neil et al.; Whitehouse Station, N.J.; Merck, 2001.), The Pill Book (9th edition; Silverman, editor; Bantam Books; 1998), Physicians' Desk Reference: PDR (Oradell, editor; Medical Economics Co., 2001) and the Martindale Drug Reference (32nd, Parfitt, editor; Pharmaceutical Pr; 1999). The mechanisms of action, half life, dosages, chemical names, and chemical structures were collected directly from the product information provided by the manufacturer or, in some cases, from the above references. Some of the chemical and physical properties of drugs were obtained from the ChemFinder web site (http://www.chemfinder.com), which is provided by CambridgeSoft ® Corporation. The synthesis information and other chemical and physical properties were extracted from the Beilstein database (Cooke and Schofield, 2001). The target protein information was collected from Therapeutic Target Database (http://xin.cz3.nus.edu.sg/ -group/ttd/ttd.asp) as well as textbooks including Principles of Medicinal Chemistry (4th Edition; Editor staff Foye et al;

Lippincott Williams & Wilkins; 1995). The metabolizing enzyme sequences were obtained from searches of the SWISS-PROT (Boeckmann et al., 2003) and NCBI GenBank (Benson et al., 2002) sequence databases. 3D structural data was obtained from the Protein Data Bank (Berman et al., 2000). These high quality data sources ensured the accuracy and quality of the data in DrugBank.

3.2.3 DrugBank Structure and Layout

DrugBank is structured in two layers. The top layer consists of the DrugBank browser pages with the bottom layer consisting of individual drug cards. Each layer is formatted in an HTML formatted table. Each of these layers is connected to a series of searching utilities written in Perl. All of the layers and searching utilities are directly accessible through the DrugBank homepage (Figure 3.1A). As can be seen from this screenshot users have 7 menu options: 1) Home; 2) Browse; 3) Structure Query; 4) Text Query; 5) Sequence Query; 6) Data Extractor and 7) Stats. This menu appears at the top of each DrugBank page, regardless of the "level" one is viewing. The Browse function displays a seven column synoptic table (a drug list) that provides short summaries of all the drugs contained in DrugBank. The information included here (Figure 3.1B) includes the DrugBank accession code, generic name, chemical formula, molecular weight, 2D structure, therapeutic category and therapeutic indication. The accession code is a 6 letter alphanumeric code containing the first three letters of the drug's generic name (a pseudo-abbreviation) followed by a three digit number. This

accession system allows up to 456 million drug compounds to be uniquely entered while at the same time allowing many experts to "approximately" identify the drug on the basis of its three letter abbreviation. The drug list can be sorted by generic name, accession code, molecular weight, CAS (Chemistry Abstract Service) number, category and indication. The sorting functions in DrugBank were implemented using Perl using a style and format similar to the PubMed sorting utilities found at the NCBI website (http://www.ncbi.nlm.nih.gov/-entrez/query.fcgi?db=PubMed). Users have the option of selecting individual drug-list pages (numbered sequentially) on the basis of page number or by advancing to the next page by pressing arrow hyperlinks. Each drug list page can be reformatted to list from 20 to 200 drugs. The intent of having a browsing utility with a variety of formatting or viewing options is to encourage users to explore or scan for information visually. Indeed, most scientists are used to viewing tables and charts and many important inferences are made by seeing a large quantity of data displayed in a compact form.

To display the results of any search or to permit simple browsing of the database itself, DrugBank uses the concept of "DrugCards" in analogy to the very popular GeneCards database (http://bioinfo.weizmann.ac.il/cards/). The leftmost column in each DrugBank drug list page contains the DrugCard link (a hyperlinked button marked DRUGCARD). The DrugCard represents the second level of information in DrugBank. As seen in Figure 3.2, it is a two column table that

Figure 3.2 A Screenshot of a Sample DrugCard in DrugBank

contains up to 60 fields with detailed information about each drug. The left column lists the field title or subject descriptors; the right column lists the actual content for the field. Some of the content fields provide hyperlinks to other databases such as the PDB (Berman et al., 2000), GenBank (Benson et al, 2002) and GeneCards (Rebhan et al., 1997). Users can visit other online databases for the relevant information by just clicking on the hyperlink. Each DrugCard contains common names, alternate names, brand names, IUPAC (International Union of Pure and Applied Chemistry) names, CAS (Chemistry Abstract Service) number, mixtures, source, manufacturer(s), MSDS (Material Safety Data Sheet) links, DIN (Drug Identification Number), 2D structural data (images), chemical formula, solubility, toxicity, physical state (solid, liquid, gas), LogP, melting/boiling point, synthesis references, 3D structure, SMILES (Simplified Molecular Input Line Entry System) string, MOL-file, PDB (Protein Data Bank) file, drug class, therapeutic indication, pharmacology, mechanism of action, drug target, prescription information, metabolites, target (protein) sequence, GenBank link (to the target protein), target structure, chromosome position (of the target protein/gene) and SNP's (single nucleotide polymorphism associated with the gene). Within each DrugCard the 3D structures of both the drugs and drug targets can be interactively viewed with hyperlinked "View" buttons. These buttons call up the embedded ACD Structure Drawing Applet. This molecular viewing applet is fully compatible with JDK 1.0.2 specifications and does not cause problems even with the older versions of popular browsers such as

Netscape Navigator 3.0. The applet was acquired at http://www.acdlabs.com/-products/java/sda/.

Each DrugCard is created dynamically using Perl CGI scripts. The CGI scripts make use of the flat files in our database and dynamically generate the HTML tags needed to display the drug information to users in a pleasantly formatted view. This dynamic formatting process makes database maintenance much easier and saves a great deal of time as one avoids having to manually generate web pages for each drug. Note that each drug in DrugBank has a fully annotated DrugCard .

3.2.4 Data Retrieval and Searching

A database should not be merely a warehouse or data. Indeed the utility of any database has to be measured by the ease and speed with which its data can be retrieved. Due consideration has to be given to the type of individuals (and skills) of the most likely users. For a drug database, different categories of users, (patients, doctors and researchers) may want to search the database in different ways. For example, patients, pharmacists or physicians may want to search the database for a particular drug or brand name. In this case, a simple text query function is needed. However, medicinal chemists or pharmaceutical researchers may want search for drugs with similar chemical structures. In this case, a structural query function is needed. Proteomics or genomics researchers may

want to query according to a sequence or polymorphism. In this case a sequence search utility (BLAST) is needed. In other cases, users may want to construct more sophisticated queries (i.e. Show me all drugs with melting points > 120 °C which are psychoactive and have molecular weights > 300 daltons). In this case a data extractor (DEXTOR) or structured query language system is needed. In order to provide all users access to a full range of search tools, the data in DrugBank was indexed using a variety of PERL scripts. Each text field was parsed into an index file (flat file format) which was subsequently accessed by a series of specially written Common Gateway Interface (CGI) scripts designed to run or access the 4 DrugBank query functions: 1) the data extractor; 2) the chemical structure search; 3) the BLAST sequence search; and 4) the Boolean text query.

3.2.5 The Data Extractor

The data extractor (DEXTOR) was designed to permit advanced or complex searches within DrugBank. Written by Melania Rouani, with modifications from Dr. Bahram Habibi-Nazhad and Dr. Anchi-Guo (both from Dr. Wishart's lab), DEXTOR was originally developed for the CyberCell database (http://redpoll.pharmacy.ualberta.ca/ -CCDB/). However, given the similarity in structure and format between DrugBank and CCDB (Sundararaj et al., submitted), it became quite apparent that DEXTOR could be adapted to work with DrugBank as well. I was partially responsible for re-indexing DrugBank to facilitate the

integration of DEXTOR with DrugBank. This utility, which was written entirely in Perl, employs a simple index-based searching system that allows users to select one or more data fields and to search for numeric ranges as well as occurrences or partial occurrences of words, strings or numbers. The interface for the data extractor uses clickable scroll-boxes (written in HTML and Java script) which contain lists of all the DrugBank field terms so that users may intuitively construct complex SQL-like queries. Each DrugBank term is linked to a pre-formatted text box that allows users to enter ranges (if it is a numeric field) or text (if it is a text field) into the selected query fields. For instance, using a few mouse clicks it is relatively simple to construct a query asking for all drugs having a melting point > 100 °C, a logP greater than 1.5 and a molecular weight less than 300 Daltons. The output from these queries is provided in a tab-delimited Excel format for easy downloading, analysis or graphical display.

In constructing the DEXTOR interface HTML frames were used to separate the scroll-box menu area and the display area. Once users select the items to extract from the scroll-box menu, Java scripts generate text entry boxes for the selected item(s) and display it (them) on the left side of the page. Once the query limits are entered and the search is submitted, the DEXTOR backend finds the indices for the selected items or items, matches all queries that fit the limits or text definitions and then dynamically generates a web page to present the search results. To make the search as efficient as possible the index files are pre-calculated using a series

of indexing scripts written in Perl. Each index file contains a specific field, such as the therapeutic indication, of every drug in DrugBank.

While it would have been possible to perform at least some of the functions described with DEXTOR using a relational database engine such as Oracle or MySQL, the cost, expertise and overhead needed to purchase, develop and maintain such a system were beyond our skills and resources. Furthermore, the size of the database was sufficiently small that no loss in performance would be expected by opting for a Perl-based search utility. Overall, the routines in DEXTOR seem to provide the speed, flexibility and customizability needed for this application. We expect DEXTOR should work well even as the database expands by up to 10 fold.

3.2.6. Structural Query

To allow users search for drugs with similar or identical structures, a simple web-based structural query tool was developed. This tool was originally written by Dr. Habibi-Nazhad and later refined by Dr. Anchi Guo and myself. Users query the database by drawing a structure using the ACD Structure Drawing Applet (http://www.acdlabs.com/products/java/sda/). This drawing applet is a Java utility that allows users to draw or select chemical substructures using their mouse or key page and to render reasonably high quality 2D chemical structures. The structure generated by a user is then transformed into a MOL file *via* the

ACD Structure Drawing Applet. Using the Babel Conversion Program (Pat Walters and Matt Stahl, 1992), which has been installed on the DrugBank server, the MOL file is transformed into a SMILES (Simplified Molecular Input Line Entry System) string, which is a linear text notation for representing 2D chemical structures (Weininger, 1988). Once a structure is converted into a text format it is relatively easy to implement Perl string search and string matching utilities to match a query structure SMILES string with SMILES structures in the DrugBank database. Specifically we developed a Perl CGI program which searches the DrugBank index file for SMILES strings (which was made in advance) and finds the drugs whose SMILES strings match or partially the query string (Figure 3.3). If the query SMILES string is found in the index file, the CGI program will dynamically generate an HTML web page displaying the structure query search results in the same format as DrugBank's drug list or drug browsing page.

3.2.7. Other Search Utilities

In addition to the advanced "relational" query system (DEXTOR) and chemical structure query tools, DrugBank also supports general text and gene/protein sequence searches. Specifically DrugBank supports a web-based local BLAST sequence search (Altschul et al., 1990) as well as a fast boolean keyword query system supported by GLIMPSE (Manber and Wu, 1994; Manber and Bigot, 1998).

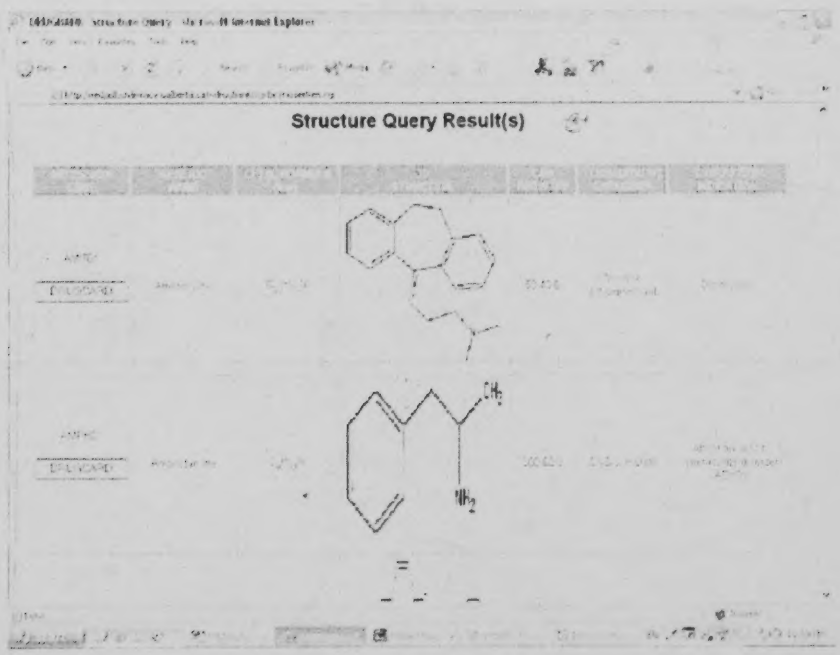


Figure 3.3 A Screen Shot of DrugBank's Structural Query Function

The BLAST search was implemented by converting all 314 sequences in DrugBank into a FASTA flatfile, installing a local version of BLAST (downloaded from NCBI) and then customizing the BLAST web interface to match the style and needs for the DrugBank Interface. The GLIMPSE search engine, which is similar to the GREP text search tool in UNIX supports Boolean (AND, OR, NOT, wild-card) multi-word text searches as well as text searches with partial matches, mis-spellings or case sensitivity. These search tools allow users to search DrugBank *via* the sequence, partial sequence, protein name, accession number, chemical name or any other keyword or combination of keywords.

One of the key strengths in DrugBank is its broad range of exact or fuzzy querying tools that can allow users, for instance, to query an entire genome (say of a newly sequenced parasite) to look for possible drug targets, or to search the database to see if a new drug lead may have possible cross reactivity to other drug targets, or alternately to determine the possible metabolic fate of a new drug lead. Using the structure query tools, newly synthesized or identified lead compounds can be compared to existing structures to assess/predict possible efficacy, cross reactivity, metabolism or physical properties. Similarly, existing drugs can be compared or analyzed for key trends, properties or features to help in drug design synthesis efforts.

3.3 Discussion

3.3.1 Data integration

DrugBank integrates a wide range of pharmaceutical, chemical and proteomic knowledge into its database. Patients can find clinical information about drugs concerning indications and dosages. Doctors and pharmacists (and pharmacy students) can learn or brush up on the pharmacology and metabolism of specific drugs. DrugBank also provides information for medicinal chemists and pharmaceutical researchers as well, including details on the synthesis, physical properties, structures and even related genomics information (SNP's, Locus and gene names) or proteomics information (protein target information, mechanism of action and the links to other important online databases such as the Protein Data Bank and SWISS-PROT).

Compared to existing, freely available web-based drug systems such as RxList (http://www.rxlist.com/), WebMD (http://www.webmd.com/) as well as the more expensive commercial drug databases such as the Martindale Drug Reference (32nd, Parfitt, editor; Pharmaceutical Pr; 1999), DrugBank appears to be unique in its coverage or linkage of small molecular information (drugs) with large molecule information (genes and proteins). Among these drug databases we have surveyed to date, only DrugBank provides 2D and 3D structures for each drug and drug target. Only DrugBank covers the metabolizing enzymes of each drug (if available). DrugBank's another unique feature is the downloadable

Material Safety Data Sheets. Furthermore, its unique web-based structural query function makes it much easier to find compounds or to identify possible structure-function relationships.

3.3.2. Database Administration and Maintenance

Ease of maintenance and facile updating were two of our main priorities in the construction of DrugBank. Based on these two concerns, all DrugBank files were formatted and stored as flat files (tab delimited text files). While frowned upon in most large database undertakings, this simple file format makes it easy to manage and distribute the information and reduces the need for software packages to maintain the database. The other advantage of working with flat files is that all the drug card web pages could be generated dynamically using relatively simple Perl scripts. This dynamism ensured that even during the input of new data or while the database is under construction, DrugBank will still accessible. To further distribute the load of maintaining and updating the database, we also allow users to edit DrugBank cards. The intent is to allow users to provide feedback and/or corrections to the database. The DrugBank editing from resembles a standard DrugCard but it has text boxes below each entry field which allow users to enter new (corrected) data. Additional comment boxes allow new fields to be added or comments to be sent to the DrugBank administrator (and acknowledgement is automatically sent to the sender). Once a DrugBank card has been edited, users may submit this information. It arrives as an email attachment in the DrugBank

administrator's mail box. From that the administrator will assess the quality and validity of information and make the changes as warranted. In this way we are able to control the content of DrugBank while at the same time correcting or adding information in response to needs or knowledge from the user community. In this way, DrugBank can continue to grow and evolve with the help of both expert users and expert administrators.

3.4. Conclusions and Future Directions

While DrugBank can and does serve a role for patients, pharmacists and doctors, we believe that many of the tools and concepts developed for DrugBank could be more directly applied to drug discovery. In this regard, we have received positive feedback from a number of researchers worldwide who have suggested that an enlarged database would better serve their purposes. Therefore, we intend to significantly increase both the size and content of the DrugBank database. In addition we intend to develop more sophisticated automated updating scripts to ensure the currency and accuracy of the data. Over the next year we will focus on expanding DrugBank's content of cancer and anti-viral drugs as well as including the current set of 100+ FDA approved protein drugs. The inclusion of protein and peptide pharmaceuticals in DrugBank will require a slight change to DrugBank's standard data fields as chemical structures and most kinds of physico-chemical data are not relevant or particularly transferable to peptides or proteins.

Given that many small molecule drugs are derived from naturally occurring metabolites or metabolic intermediates (amino acids, nucleotides, vitamins, etc.) we believe that the utility of DrugBank could be increased by including these compounds in the database as well. An excellent source for this information will be the CyberCell Database (CCDB – http://www.ccdb.ualberta.ca). The CCDB format (Sundararaj et al., submitted) developed for storing and displaying metabolite information closely parallels the structure used for DrugBank. Therefore we expect it will take very little effort to eventually incorporate this metabolite data into DrugBank.

In addition to expanding the database content, we also intend to extend DrugBanks search capabilities. The chemical structure search utilities are still somewhat rudimentary and could be improved with the used of Ullmann's substructure methods (Ullmann, 1976). Likewise, the support for full proteome searches (a query proteome with 3000+ proteins) against the database still needs to be improved. This proteome search utility will be critical for allowing researchers to quickly identify possible protein targets from newly sequenced pathogens or viruses. The implementation of a docking utility with automated free energy evaluation (Structure Thermodynamic Calculator – Osterberg et al., 2002) is also planned.

Throughout these updates and developments we intend to keep DrugBank as public and as freely available as possible. We intend to release the source code for most of the key programs and CGI scripts used in DrugBank to encourage others to develop or improve on the code and designs. Likewise the data will also be made downloadable so that others may use the information in developing more specialized databases or applications. In making much of DrugBank freely available over the web we are hoping that this database (and this project) will evolve towards becoming a "community project", allowing independent submissions, corrections, development and additions. Such community projects as GenBank and the PDB have grown to be very successful and important resources in molecular and structural biology. Hopefully the same may happen with DrugBank as it relates to pharmaceutical science.

Chapter 4; Expression, Purification and Preliminary Characterization of an Unknown Protein - M.t.0776

4.1 Introduction

4.1.1 *Methanobacterium thermoautotrophicum* (M.t.) and the M.t.0776 protein

In this chapter, I will describe my efforts in trying to express and characterize a protein from *Methanobacterium thermoautotrophicum* (M.t.) ΔH. *M. thermoauotrophicum*. is an archaebacterium originally isolated in 1971 from a sewage pond in Urbana, IL. (Zeikus et al., 1972). The reason why the name has a prefix "methano-" is that the organism can synthesize methane (CH_4) by reducing carbon dioxide (CO_2) using hydrogen (H_2). The prefix "thermo-" means that the organism grows at high temperatures. Actually, it grows at temperatures ranging from 40 to 70°C (optimally at 65°C). The complete genome sequence of *Methanobacterium thermoautotrophicum* ΔH was determined and publicly released in 1997 (Smith et al., 1997). Three years later, this relatively obscure bacterium was chosen by the Ontario Cancer Institute (OCI) as a prototype organism with which structural proteomics studies would be conducted. *M. thermautotrophicum* was chosen for 4 reasons: (1). The complete genome sequence of *M. thermautotrophicum* was known. (2). The genome does not include introns, making it less expensive to obtain clones. (3). *M. thermautotrophicum* grows at high temperatures, making its proteins stable enough to routinely endure the purification process. (4). The high GC content of

the *M. thermautotrophicum* DNA proved to be a good substrate for PCR cloning (Christendat et al., 2000).

Of the 1871 potential open reading frames (ORFs) in *M. thermautotrophicum*, we chose one particular clone labeled as M.t. 0776. This gene codes for a 101 residue protein of unknown structure and function. The protein's DNA and amino acid sequence are shown in Figure 4.1. The molecular weight of the protein is predicted to be 11.397 kDa and its estimated pI is 5.42, as calculated using the Compute pI/Mw tool found at the ExPASy website (Bjellqvist et al., 1993). A BLAST search (Altschul et al., 1990) conducted on September 1st, 2003 showed that only a small number of conserved or hypothetical proteins exhibit some similarities with M.t.0776 (Figure 4.2). Since M.t.0776 has little sequence identity with other proteins of known structure or function, it was assumed that this protein may have a novel 3D fold.

As mentioned in Chapter 1, one of the key objectives of structural proteomics is to identify and characterize proteins with novel 3D structures. The postulated novel fold for M.t.0776 may serve as a structural template for modeling or understanding the role or function of a new class of proteins. Furthermore, since this protein is bacterial in origin, it could be similar to other essential proteins found in, as yet undiscovered, bacterial pathogens. As such, this kind of structural information may be useful in the development or discovery of potential drugs or

The DNA sequence of M.t. 0776:

ATGACGTTCTGCCTGGAGACCTATCTGCAGCAGTCAGGGGAATATGAGATCCACA

TGAAGAGGGCCGGTTTCAGGGAATGTGCAGCAATGATAGAAAAGAAGGCACGGAG

GGTCGTCCACATAAAGCCCGGTGAGAAGATCCTCGGCGCAAGGATAATCGGCATA

CCACCCGTACCCATCGGTATAGATGAGGAAAGATCAACCGTCATGATACCATACA

CCAAGCCATGCTACGGTACAGCCGTGGTTGAGCTCCCGGTTGATCCAGAGGAGAT

AGAGAGGATCCTTGAGGTGGCAGAACCCTGA

The protein sequence of M.t. 0776

MTFCLETYLQQSGEYEIHMKRAGFRECAAMIEKKARRVVHIKPGEKILGARIIGI

PPVPIGIDEERSTVMIPYTKPCYGTAVVELPVDPEEIERILEVAEPz

The sequence of expressed protein with 6xHis tag and thrombin cleavage site

(MGSSHHHHHHSSGLVPRGSH)MTFCLETYLQQSGEYEIHMKRAGFRECAAMIEK

KARRVVHIKPGEKILGARIIGIPPVPIGIDEERSTVMIPYTKPCYGTAVVELPVD

PEEIERILEVAEP

Figure 4.1 The DNA and Protein Sequence of M.t.0776

```
Db AC        Description                                          Score E-value

tr Q26870    Conserved protein [MTH776] [Methanobacterium therm... 246 3e-64
tr Q8TY94    Uncharacterized protein conserved in archaea [MK04... 45 4e-12
sp Q57895    Y453_METJA Hypothetical protein MJ0453 [MJ0453] [M... 42 2e-09
tr Q8TUD8    Hypothetical protein MA0129 [MA0129] [Methanosarci... 33 7e-08
tr Q8PX07    Conserved protein [MM1415] [Methanosarcina mazei (... 34 1e-07
tr Q8TUE1    Hypothetical protein MA0126 [MA0126] [Methanosarci... 37 1e-05
tr Q9WTK7    LKB1 [STK11] [Mus musculus (Mouse)]                  32  6.0
tn AAH52379  Hypothetical protein [Mus musculus (Mouse)]          32  6.0
tn CAD71497  Ammonium transporter [amtB] [Pirellula sp]           32  6.0
tr Q9R0P1    LKB1 serine/threonin kinase (Fragment) [STK11] [Mu... 32  6.0


tr Q26870 Conserved protein [MTH776] [Methanobacterium thermoautotrophicum] 101 AA
                                                                            align

 Score = 246 bits (526), Expect = 3e-64
 Identities = 101/101 (100%), Positives = 101/101 (100%)

Query: 1   MTFCLETYLQQSGEYEIHMKRAGFRECAAMIEKKARRVVHIKPGEKILGARIIGIPPVPI 60
            MTFCLETYLQQSGEYEIHMKRAGFRECAAMIEKKARRVVHIKPGEKILGARIIGIPPVPI
Sbjct: 1   MTFCLETYLQQSGEYEIHMKRAGFRECAAMIEKKARRVVHIKPGEKILGARIIGIPPVPI 60

Query: 61  GIDEERSTVHIPYTKPCYGTAVVELPVDPEEIERILEVAEP 101
           GIDEERSTVHIPYTKPCYGTAVVELPVDPEEIERILEVAEP
Sbjct: 61  GIDEERSTVHIPYTKPCYGTAVVELPVDPEEIERILEVAEP 101


tr Q8TY94 Uncharacterized protein conserved in archaea [MK0411] 97 AA
          [Methanopyrus kandleri]                                align

 Score = 45.4 bits (92), Expect(3) = 4e-12
 Identities = 19/42 (45%), Positives = 26/42 (61%)

Query: 7  TYLQQSGEYEIHMKRAGFRECAAMIEKKARRVVHIKPGEKIL 48
           T+  + EYEI M R  F +CA  IE K   +V +++PGE+IL
Sbjct: 2  TFCLEEREYEILMARRPFDDCARYIESKFGNIVKLQPGEEIL 43


 Score = 38.4 bits (77), Expect(3) = 4e-12
 Identities = 13/29 (44%), Positives = 22/29 (75%)

Query: 69 VHIPYTKPCYGTAVVELPVDPEEIERILE 97
           +++P T+PC+G+ VV++ V  EE+E  L+
Sbjct: 63 IVLPITKPCHGSFVVKIEVSAEELEWFLK 91


 Score = 18.6 bits (34), Expect(3) = 4e-12
 Identities = 5/5 (100%), Positives = 5/5 (100%)

Query: 1 MTFCL 5
           MTFCL
Sbjct: 1 MTFCL 5


sp Q57895      Hypothetical protein MJ0453 [MJ0453] [Methanococcus jannaschii] 107 AA
   Y453_METJA                                                                 align

 Score = 41.7 bits (84), Expect(3) = 2e-09
 Identities = 15/30 (50%), Positives = 22/30 (73%)

Query: 59 PIGIDEERSTVHIPYTKPCYGTAVVELPVD 88
           PI + E + V+ PYTKPCYGT V+++ +D
Sbjct: 51 PIPVAYEDNYVIFPYTKPCYGTFVLKINLD 80


 Score = 26.9 bits (52), Expect(3) = 2e-09
 Identities = 9/27 (33%), Positives = 16/27 (59%)

Query: 21 RAGFRECAAMIEKKARRVVHIKPGEKI 47
           + GF+ECA  I K  + +  + G +I
Sbjct: 14 KGGFKECAEYIRKNFKNIKEMEAGYEI 40


 Score = 25.0 bits (48), Expect(3) = 2e-09
 Identities = 7/16 (43%), Positives = 12/16 (75%)

Query: 49 GARIIGIPPVPIGIDE 64
           G  +GIPP+P+  +E
Sbjct: 43 GIFLIGIPPIPVAYED 58
```

Figure 4.2 The BLAST Search Result of M.t.0776 by NCBI BLASTP 2.2.5

Conducted on September 1, 2003

drug targets.

4.1.2 Affinity chromatography in protein purification

In order to make NMR samples of M.t. 0776, the protein needed to be isolated and purified. There are many methods available to separate proteins according to their different chemical and physical properties, including electrophoresis, centrifugation, solvent precipitation, crystallization and chromatography. Chromatography is one of the most frequently used methods for protein purification. Common chromatographic methods include ion-exchange, hydrophobic interaction, gel-filtration, and affinity-based techniques. Among them, affinity chromatography is one of the most powerful protein purification techniques available.

In this particular study, we adopted the poly His tag technology of QIAGEN™ to purify M.t. 0776 using affinity chromatography. Specifically, our M.t. 0776 clone was designed to contain a 6 residue poly His tag at N-terminus of the M.t. 0776 protein. The 6xHis affinity tag facilitates binding to a Ni-NTA (Nickel nitrilotriacetic acid) column. Using QIAGEN™'s pQE vectors (Bujarg et al., 1987; Stuber et al., 1990; Schwarz et al., 1978) the histidine tag can be placed at the C- or N- terminus of the protein of interest. It is poorly immunogenic, and at pH 8.0 the tag is small and uncharged. Therefore the His tag does not generally affect secretion, compartmentalization, or folding of the fusion protein within the

cell (The QIAexpressionist, 2001). The His tag is an affinity tag that binds strongly to Nickel-NTA under appropriate salt and pH conditions.

NTA is a tetradentate chelating adsorbent originally developed at Hoffmann-La Roche. It overcomes the low yields obtained by using traditional chelating ligand - iminodiacetic acid (IDA) methods (Sulkowski, 1985). With NTA, four of the six ligand binding sites are occupied by the nickel ion, leaving two sites free to bind with the 6xHis tag (Figure 4.3). In this way, the NTA resin can bind 6xHis-tagged proteins, making it possible to purify proteins from less than 1% of the total protein preparation to greater than 95% homogeneity in a single step (Janknecht et al., 1991).

4.2 Materials and Methods

4.2.1. Materials

Double distilled water (ddH_2O) was prepared using a Corning AG-11 generator and Corning 4R9 automatic collection system (Corning, NY, USA). Potassium dihydrogen orthophosphate, sodium dihydrogen orthophosphate, sodium dodecyl sulphate, D-glucose, ammonium chloride, and magnesium sulphate were purchased from BDH Inc. (Toronto, ON). Tris aminomethane, sodium chloride, and sodium phosphate (dibasic) were obtained from Caledon Laboratories Ltd. (Georgetown, ON). Lysozyme, imidazole, and ampicillin were purchased from Sigma Chemical Co. (St. Louis, MO, USA). Tryptone Peptone, Yeast Extract and

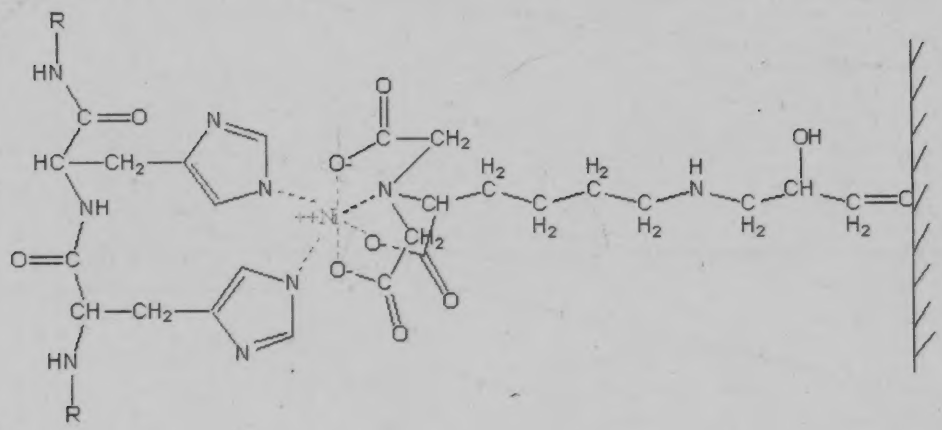


Figure 4.3 Interactions between Neighboring Residues in the 6xHis Tag and Ni-NTA Matrix

Bacto Agar were obtained from Becton Dickinson Company (Sparks, MD, USA). Isopropyl β-D-thiogalactopyranoside (IPTG) was purchased from Rose Scientific Ltd. (Edmonton, AB). Ni-NTA Agarose was obtained from QIAGEN (Gmbh, D-40724 Hilden, Germany). UV absorbance was measured on Ultrospec 3000 UV/Visible Spectrophotometer from Pharmacia Biotech (Peapack, New Jersey, USA), NMR spectra were acquired on a Varian 500 MHz Inova spectrometer (Palo Alto, CA, USA).

4.2.2. Transformation

M.t. 0776 was cloned into an *Escherichia coli* expression vector, pET15b vector (Novagen), by our collaborators in the Ontario Cancer Institute (OCI) as part of their structural proteomics screening efforts (Christendat et al., 2000). The clone contains an N-terminal 6xHis tag followed by a thrombin cleavage site (Christendat et al., 2000). Two primers were used to obtain the PCR clones of M.t.0776. One PCR (Figure 4.4) primer contains the 5' restriction site (*Eco*RI, R1), the 5'-CAT CAC CAT CAC CAT CAC-3' sequence for the 6xHis tag, and a priming sequence that is homologous to the 5' end of the M.t.0776 ORF corresponding to the N-terminus of the protein. The other primer contains another restriction site (*Nco*I, R2) which is located within the ORF. The screened PCR clones and the original vector were then both cut with the enzymes *Eco*RI and *Nco*I and ligated together.

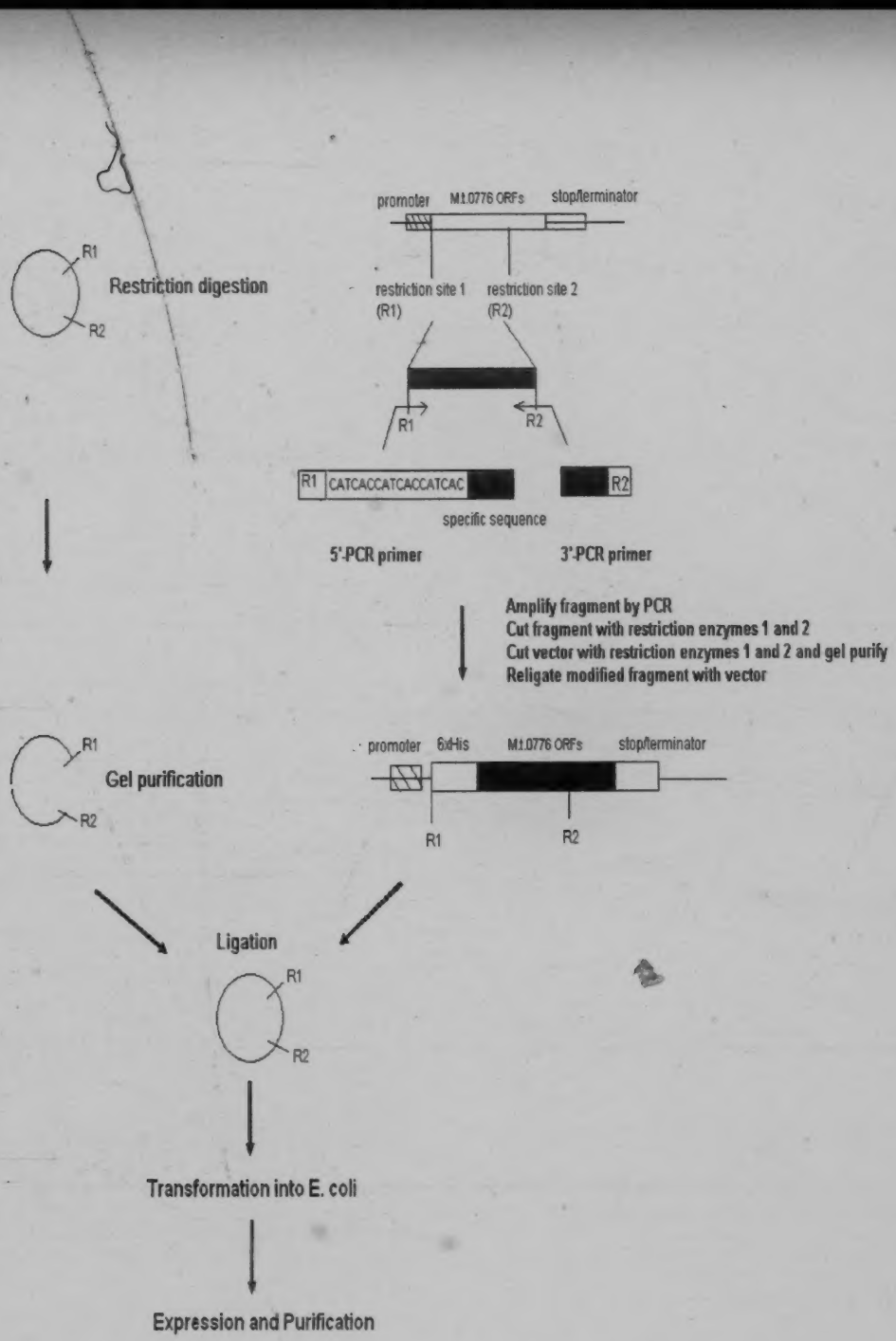


Figure 4.4 Construction of the M.t.0776 Expression Vector

The plasmid was transformed into the Epicurian *E. coli* BL21 (DE3) strain provided by Stratagene (La Jolla, CA, USA). This strain was used as a host for both expression of M.t.0776 and plasmid propagation. The plasmid was transformed into competent cells through heat-pulsing according to manufacturer's instructions (Stratagene).

4.2.3. Expression

A single colony was selected from an LB + amp (Luria broth + Ampicillin) plate grown overnight. The colony and was put into 25 mL LB media (or M9 to test labelling efforts) containing 7.5% ampicillin. The culture was incubated overnight at 31°C with shaking at 320 rpm. The next day, 20 mL of the overnight culture was added to 4 flasks containing 250 mL LB or M9 media containing 7.5% ampicillin. The culture was incubated at 31°C with shaking at 320 rpm. The OD value was measured at 600 nm and cells were induced with 1 mM IPTG when the OD_{600} reached 1.2. Before the cells were induced, a 1 mL sample of culture was taken out and centrifuged at 13000 rpm for 5 minutes; this pellet was saved for later SDS-PAGE analysis. To induce the cells, IPTG was added into each flask to a final concentration of 1 mM and the culture incubated at 37°C with shaking at 320 rpm for 6 hours. Every 2 hours, 1 mL of culture was removed and centrifuged at 13000 rpm for 5 minutes and the pellets saved for later SDS-PAGE analysis. After the induction period was complete, the culture was divided into 2 bottles and centrifuged for 1.5 hours at 3000 rpm, with the pellets being saved for later

extraction. Lysis buffer (25 mL; 50 mM NaH_2PO_4, 300 mM NaCl, 10 mM imidazole, pH=8.0) was added to the pellet. To release the protein from the cellular cytoplasmic fraction, the buffer-pellet mixture was freeze-thawed 3-4 times, then centrifuged at 12000 rpm for 1.5 hours. The pellet and supernatant were saved for later SDS-PAGE gel analysis and subsequent processing.

4.2.4. Purification

Ni-NTA agarose (12-13 mL) was loaded into a 1 cm x 20 cm column and washed with lysis buffer until a constant UV absorbance baseline (measured at 260 nm) was obtained. The cell supernatant was loaded at a flow rate of 1.5 mL/minute, and the column was again washed with lysis buffer. The column was then washed with two wash buffers (20 and 80 mM imidazole in 50 mM NaH_2PO_4, 300 mM NaCl, pH = 8.0) at a flow rate of 1.5 mL/minute. The flow-through was collected for SDS-PAGE analysis. After the washing steps, the column was eluted with two elution buffers (150 mM and 250 mM imidazole, 50 mM NaH_2PO_4, 300 mM NaCl, pH=8.0) at a flow rate of 1.5 ml/minute. The eluent fractions were collected for later SDS-PAGE analysis.

4.2.5. Mass Spectrometric Analysis of the Gene Product

In order to determine if the expression and purification were successful, selected gel bands from the SDS-PAGE of the Ni-NTA purified protein were analyzed by electrospray MS. Prior to MS analysis, the gel band were trypsinized. Specifically

individual bands were excised from the gel (using an Exacto knife) and put into a 1.5 mL centrifuge tube (Rose Scientific). Approximately 100 µL of 25 mM NH_4HCO_3 /50% acetonitrile were added to the acrylamide gel slab. The tube was then vortexed for 35-40 min on a low setting and the supernatant discarded. This wash/dehydration step was repeated ~3 times. The gel was then dehydrated with 100 µL acetonitrile. The gel particles were rehydrated in 25 µL trypsin solution and placed on ice for 10-15 minutes. The excess trypsin solution was removed and the rehydrated gel particles were overlayed with 30 µL of 25mM NH_4HCO_3 to keep them immersed throughout digestion. The solution was then incubated for 12 to 16 hrs at 37°C. After incubation, 5 µL of 5% aqueous trifluoroacedic acid (TFA) was added to halt the digestion process. Mass spectrometry was performed on a Quadrupole Time-of-Flight (Q-TOF) mass spectrometer (Bruker Reflex) by Lorne Burke, Institution of Biomolecule design, University of Alberta. The fragmentation spectra of the tryptic peptides of the protein were collected and the peptides were identified using the MS analysis program called Mascot (Perkins et al., 1999).

4.2.6. NMR Sample Preparation and Spectra Collection

Prior to sample concentration, a centricon (Millipore) with a 10 kDa molecular weight cut-off was washed with 3x10mL ddH_2O to remove glycerol, *via* centrifugation at 3000 rpm for 10 minutes. To concentrate the M.t. 0776 sample, the pooled fractions from the NTA-affinity purification process (~8 mL) were

added to the now-clean centricon filter and the sample centrifuged at 3000 rpm (4°C), until the solution volume was concentrated to ~1000 µL. Because the imidazole in the protein solution could interfere with the protein NMR signals, about 15 mL of a pure salt buffer (50 mM NaH_2PO_4, 150 mM NaCl, pH=7.5) was added to the top of the centricon apparatus, and the sample was centrifuged at 3000 rpm at 4°C until the volume of the solution reached ~1000 µL. This process was repeated for 4 times. In the final step, the solution was concentrated to 500 µL and transferred to a sterile Eppendorf tube. An internal chemical shift standard (DSS, 12 µL, 0.1 mM final concentration) and 12 µL of 3% NaN_3 (aq) were added to the protein solution. The resulting solution was transferred into an NMR tube, centrifuged for 5 minutes to remove foam and NMR spectra were collected. NMR experiments on M.t.0776 were conducted on a 500 MHz Varian Inova NMR spectrometer equipped with a 5 mm triple resonance probe. 1D ^1H spectra were collected at 25°C using 512 scans with an acquisition period of 1.4 s and a relaxation delay of 1.0 s. A 9.3 us 90° pulse was used. Spectra were baseline corrected and processed using a 0.5 Hz line broadening function.

4.2.7 Bioinformatic Analysis

One of the principal aims of structural proteomics research is to reveal the function and to determine the structure of the unknown protein by any means possible (experimental or theoretical). In this regard, bioinformatics methods developed over the past decade for predicting secondary and tertiary structure,

location, active sites and other possible features can be particularly helpful. Obviously any predicted function, location or structure would still need to be confirmed by further experiments, these predictions can certainly be used as a guide to conduct further tests or undertake additional experiments. In other words, bioinformatics predictions can narrow down the scope of our study and direct the course of further work.

In this study, the secondary structure of M.t. 0776 was predicted by NNPREDICT on ExPASy (http://www.cmpharm.ucsf.edu/~nomi/-nnpredict.html). The web site ProtScale (http://us.expasy.org/cgi-bin/protscale.pl) on ExPASy was used to analyze the protein's hydrophobicity distribution. 3D structure predictions of M.t. 0776 were conducted on the 3D-PSSM web server (http://www.sbg.bio.ic.ac.uk/~3dpssm/).

4.3 Results and Discussions

4.3.1 Transformation, Expression and Purification

The initial transformation was found to be successful, as the SDS-PAGE results showed a significant level of expression of the target protein (Figure 4.4). As expected, the intensity of one of the protein bands increased linearly with the induction time. Compared with the position of our molecular weight standard, lysozyme (with a molecular weight of 14.3 kDa), the band of interest had a similar but a little larger molecular weight. This result corresponds well with the

predicted molecular weight for M.t. 0776 with a 6xHis tag and a thrombin cleavage poly-linker (14.6 kDa). It was observed that the yields fell over time, because the antibiotic that is used, ampicillin, is only stable for 5-6 hours. This problem was solved by adding a second dose of the antibiotic at induction time to ensure successful expression, as shown in Figure 4.5.

Since the target protein has a 6xHis tag, only this protein has the ability to bind to the NTA agarose column matrix. Numerous fractions, including the wash buffer, flow through and suspected protein fractions were collected. All fractions were analyzed by SDS-PAGE. As seen in Figure 4.6, the suspected protein fractions showed the presence of a pure protein band with a molecular weight of ~14.6 kDa (Figure 4.6). During the processing of numerous M.t.0776 samples, we found that the protein was quite sensitive to freezing and freeze-drying. In particular, frozen samples, upon thawing frequently led to significant precipitation. Consequently all M.t. 0776 samples were stored at 4°C to prevent freezing and to reduce the possibility of precipitation. Similarly, samples were concentrated using Centricon filters as opposed to being concentrated by freeze drying. It was also found that a significant portion of the M.t. 0776 protein could be recovered from the cell pellet, indicating that the protein was probably forming inclusion bodies and precipitating in the cell. Similarly, samples were concentrated using Centricon filters as opposed to being concentrated by freeze drying. It was also found that a significant portion of the M.t. 0776 protein

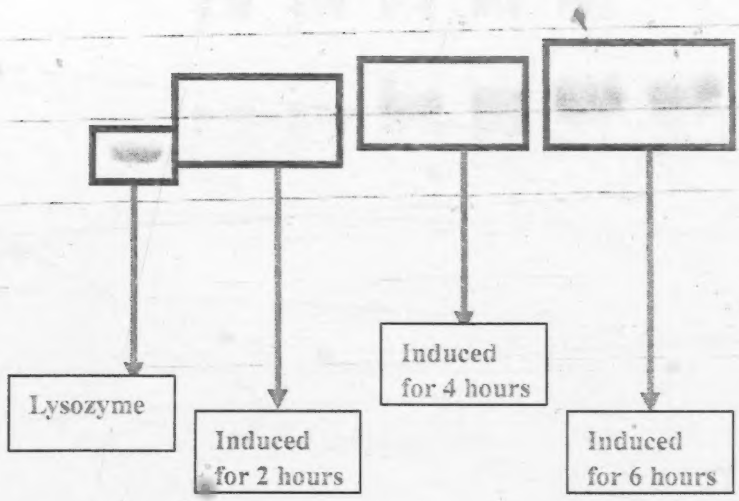


Figure 4.5 1D SDS-PAGE Gel of M.t. 0776 Expression Levels

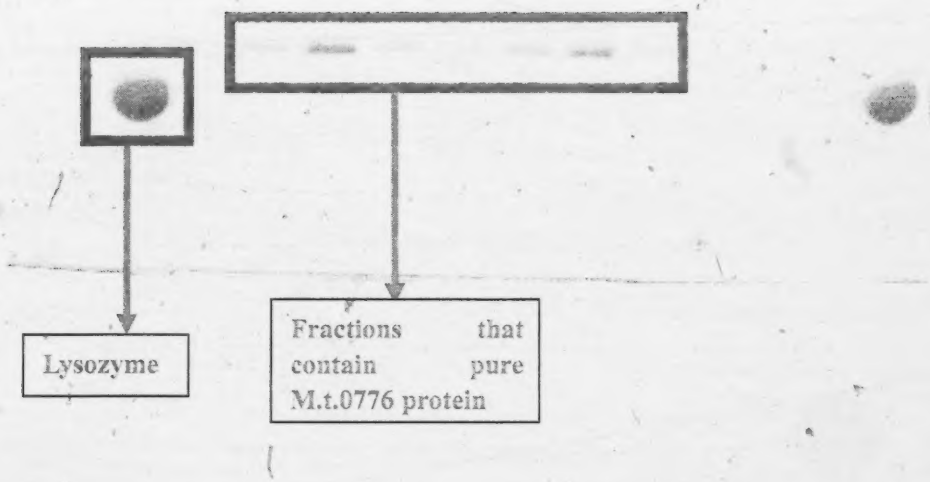


Figure 4.6 SDS-PAGE Gel of the Purification Result of M.t.0776

could be recovered from the cell pellet, indicating that the protein was probably forming inclusion bodies and precipitating in the cell. including M.t. 0776, precipitated out. A probable reason is that the protein may be well protected by the conditions of the cytoplasm; for example, the high salt concentrations and near-neutral pH in the cytoplasm may help to stabilize the protein.

4.3.2 MS and NMR Analysis

After the protein was purified, a mass spectrometric analysis was conducted at the IBD (the Institute for Biomolecular Design, University of Alberta). The result (Figure 4.7) proved that the band in the SDS-PAGE gel was, in fact, the M.t. 0776 protein. Protein yield was determined using the bovine serum albumin (BSA) assay (Bradford, 1976). The total amount of the protein expressed was 29.4 mg from 2L minimal media. A portion of this protein sample was used to collected a 1D ^{1}H NMR spectra was collected (Figure 4.8).

The rate limiting step in this study, as with almost all structural proteomics studies was the production and purification of pure protein samples that exhibit good stability in an NMR tube. Obviously, the quality of the NMR sample will have direct impact on the result of NMR analysis. We can check the quality of our protein sample by looking at two regions of the NMR spectrum. The presence of

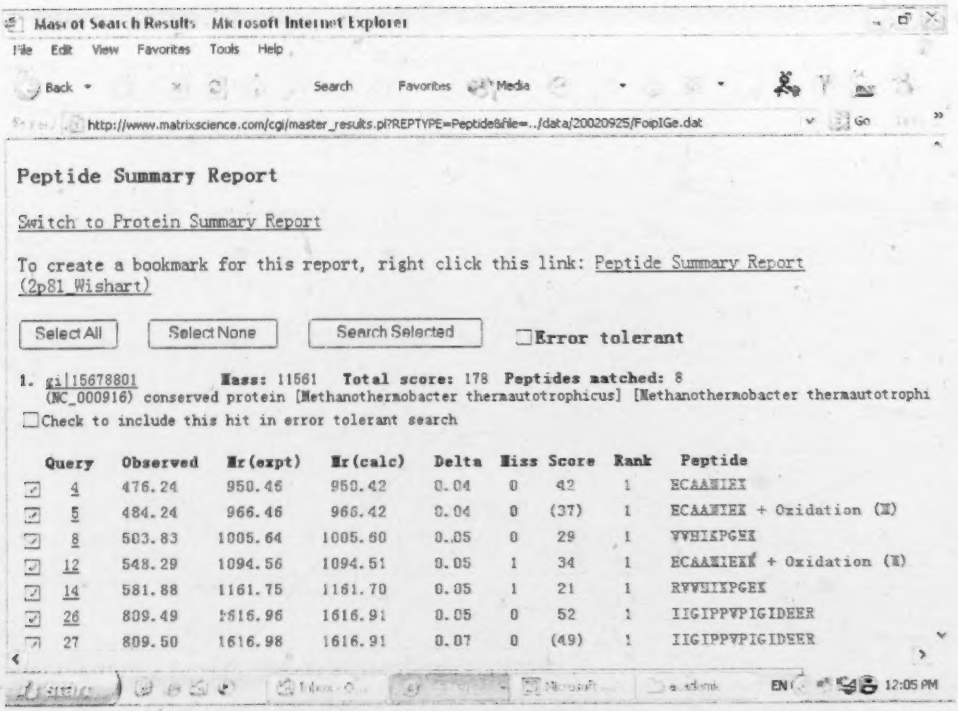

Mascot Search Results - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
http://www.matrixscience.com/cgi/master_results.pl?REPTYPE=Peptide&file=../data/20020925/FoipIGe.dat
Peptide Summary Report
Switch to Protein Summary Report
To create a bookmark for this report, right click this link: Peptide Summary Report (2p81_Wishart)
Select All
Select None
Search Selected
Error tolerant
1. gi|15678801 Mass: 11561 Total score: 178 Peptides matched: 8
(NC_000916) conserved protein [Methanothermobacter thermautotrophicus] [Methanothermobacter thermautotrophi
Check to include this hit in error tolerant search
Query Observed Mr(expt) Mr(calc) Delta Miss Score Rank Peptide
4 476.24 950.46 950.42 0.04 0 42 1 ECAAMIEK
5 484.24 966.46 966.42 0.04 0 (37) 1 ECAAMIEK + Oxidation (M)
8 503.83 1005.64 1005.60 0.05 0 29 1 VVHIKPGMK
12 548.29 1094.56 1094.51 0.05 1 34 1 ECAAMIEKK + Oxidation (M)
14 581.88 1161.75 1161.70 0.05 1 21 1 RVVHIKPGEK
26 809.49 1616.96 1616.91 0.05 0 52 1 IIGIPPVPIGIDEER
27 809.50 1616.98 1616.91 0.07 0 (49) 1 IIGIPPVPIGIDEER
EN 12:05 PM

Figure 4.7 Mass Spectroscopy Analysis by IBD of an SDS-PAGE Gel in the Purification of M.t. 0776

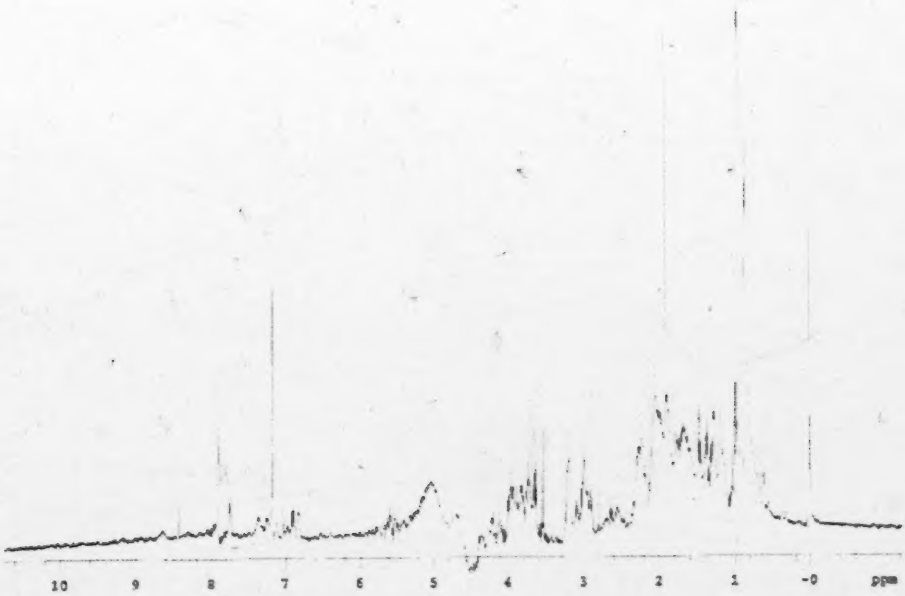


Figure 4.8 1D ^{1}H NMR Trace of M.t. 0776 Acquired at 500 MHz

signals between 0 to 1 ppm and a good dispersion of resonances between 7.3 to 9.5 ppm are evidence that the protein is in its native conformation.

If the protein is in its native conformation, the magnetic shielding of the methyl or amide hydrogen will cause the peaks to shift away from their random coil values (Wishart et al., 1994), which results in different methyl and amide groups having significantly different chemical shifts than the random coil values. As shown in Figure 4.8, the 1D ^{1}H NMR spectrum of M.t.0776 shows a number of signals between 0 to 1 ppm and good spectral dispersion between 7.3 to 9.5 ppm, which suggest that the protein is properly folded. Subsequent studies by my colleague, Godwin Amegbey, have identified much more appropriate solution conditions (yielding higher quality NMR spectra) and the protein has now been fully assigned. A 3D structure should be completed shortly.

4.3.3 Bioinformatic Analysis of M.t. 0776 Protein

Though M.t.0776 shows little sequence identity to other proteins of known function or structure, we can still predict certain aspects of this protein's using bioinformatics tools, such as the secondary structure prediction, charge and hydrophobicity distributions, and 3D structure prediction.

4.3.3.1 *Secondary Structure*

Prediction of M.t. 0776's secondary structure can help determine its overall 3D structure. There are several software and internet tools that can accurately (>75%) predict a protein's secondary structure (Rost, 2001) – many of which are found at ExPASy. Using the NNPREDICT web server on ExPASy and the sequence of M.t. 0776 protein, we predicted the secondary structure. NNPREDICT is an example of a neural network prediction method (http://www.cmpharm.ucsf.edu/~nomi/nnpredict.html). A number of web servers predict secondary structure using "Neural Network" (Bishop, 1996) approaches. In most cases, the accuracy of these prediction methods approaches 75%. Under ideal cases, the prediction accuracy can approach 79% (for all-alpha proteins). Figure 4.9 shows the predicted secondary structure determined for the M.t.0776 protein using the NNPREDICT server (http://www.cmpharm.ucsf.edu/~nomi/nnpredict.html).

As the above result shows, there are still some intermediate or difficult-to-classify regions, which could potentially exist as random coils or unstructured loops. The protein shows a mix of α-helices and β-strands (21.57% of α-helices and 36.27% of β-strands) indicating that this protein likely falls into the "mixed alpha/beta class" of proteins. Inspection of the predicted helices and beta strands suggests that the helices likely surround a core of short beta strands which coalesce into a 3-4 stranded sheet.

```
MTFCLETYLQ QSGEYEIHMK RAGFRECAAM

---EHHHH-- ----HHHHHH ----HHHHHH

IEKKARRVVH IKPGEKILGA RIIGIPPVPI

HHHHHHHEEE -------E-- EE--------

GIDEERSTVM IPYTKPCYGT AVVELPVDPE

--------EE ---------- -EEE------

EIERILEVAE P

HHHHHHHH-- -
```

(H = helix, E = strand, - = no prediction)

Figure 4.9 Prediction of M.t. 0776's Secondary Structure by NNPREDICT

4.3.3.2 *Hydrophobicity Distribution*

Hydrophobicity is one of the most dominant factors affecting the folding patterns of proteins. Hydrophobic residues are more likely to be found inside the protein (in a hydrophobic environment) than are hydrophilic residues. Likewise hydrophilic residues are frequently found on the exterior of proteins compared to hydrophobic residues. Bioinformatics tools can help us determine the hydrophobic distribution of amino acids in proteins. There are many algorithms to calculate protein hydrophobicity, such as Kyte and Doolittle's method (Kyte J. and Doolittle, 1982), Abraham and Leo's method (Abraham and Leo, 1987) as well as Bull and Breese's method (Bull and Breese, 1974). The Kyte-Doolittle algorithm calculates an average hydropathy value using a sliding window method for each position in a given sequence. Specifically, for each position, the algorithm considers the hydropathy index of the 2 neighboring amino acids as well as the central residue, and then calculates their average. This average value is the one plotted on the graph for the central position. Because of the widespread use of the Kyte-Doolittle algorithm and hydrophobicity parameters, we selected it to analyze the M.t. 0776 protein (Figure 4.10).

As can be seen from Figure 4.9, the hydrophobicity distribution roughly corresponds to the predicted locations of key secondary structure elements. In particular, the most hydrophobic regions correspond to the location of the predicted β-sheets. This observation suggests that the residues between 45 to 65

may be folded inside the protein and likely form the bulk of the protein's β-sheet. Meanwhile the residues between 4 to 20 are probably on the outside of the protein and should fold into an α-helix. These preliminary predictions can further facilitate the prediction of the 3D structure of M.t.0776.

4.3.3.3 *Preliminary 3D Structure Prediction*

Using the 3D-PSSM (Kelley et al., 1999) web server, we attempted to predict the 3D structure of M.T. 0776. Figure 4.11 shows three of the 3D-PSSM-predicted structures of M.t. 0776. 3D-PSSM uses a "threading" method, which aligns the target sequence with other similar, previously solved protein sequences. First, the server takes the query sequence and searches through its library of 1D protein sequence profiles to find similar sequences or fragments. Then, the server performs a superposition using the SAP program (Orengo et al., 1992; Taylor and Orengo, 1989). Only structures that superimpose with a weighted root mean square deviation < 6.0Å to the master structure are considered. A three-state secondary structure assignment (coil, helix, and strand) is then made on a per residue basis using STRIDE (Frishman & Argos, 1995). Then the server assigns a score, with lower scores corresponding to better predictions.

Since M.t.0776 has few sequence homologues, the prediction provided by 3D-PSSM may not be accurate. Among the three predictions shown in Figure

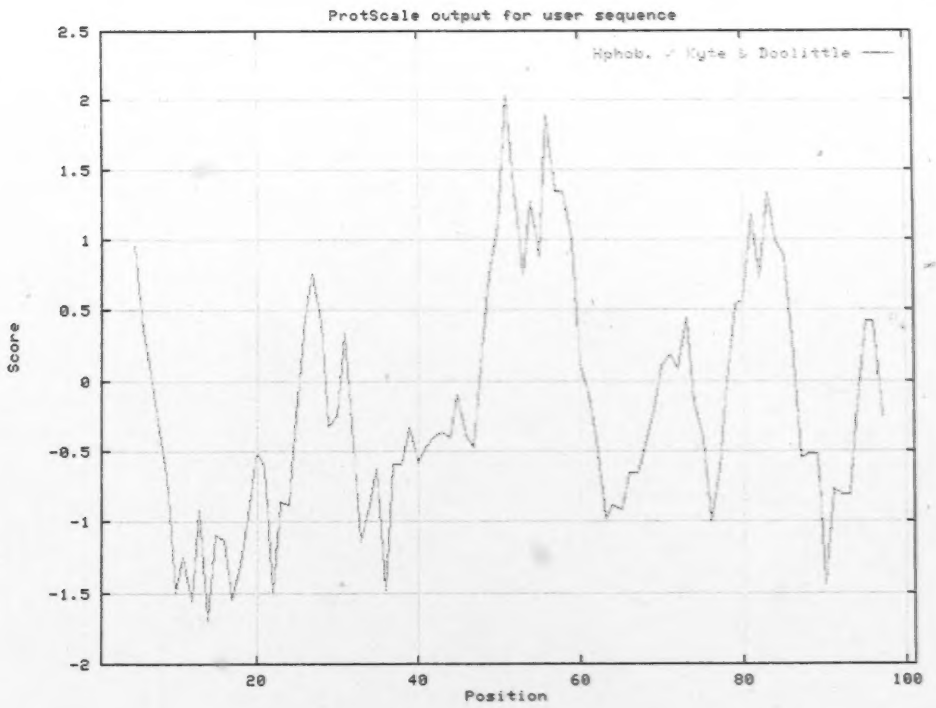
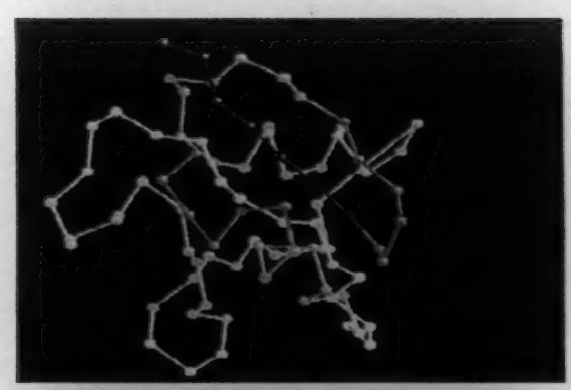
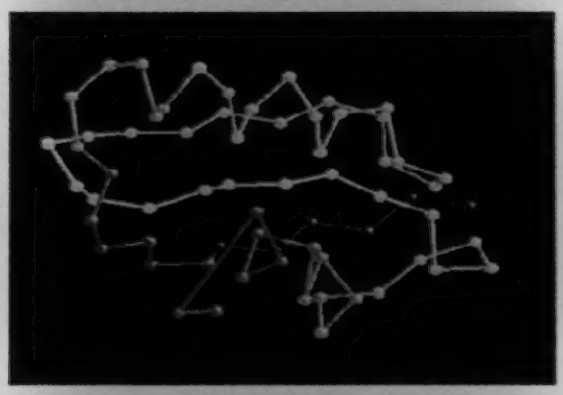


Figure 4.10 Hydrophobicity Distribution Analysis of M.t. 0776 by ProtScale



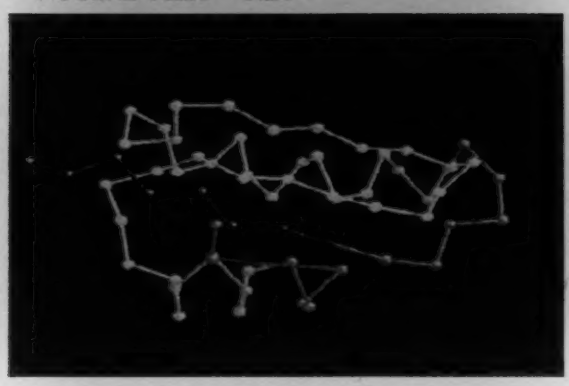
Solution Structure of the Ras-binding Domain of the Protein2 Kinase byr2 from ***S. pombe***

PSSM E-value = 6.98



Ferredoxin -like Metal -binding Domain of Menkes copper-transportin g ATPase

PSSM E-value = 9.24



Ferredoxin-like Metal-binding Domain of ATX1 Metallochap -erone Protein

PSSM E-value = 10.20

Figure 4.11 Possible Structures of M.t. 0776 Predicted by 3D-PSSM (E-value is the accuracy estimation of the prediction. The larger the value is, the less reliable the prediction.)

```
F:---EEEEEEE ----HHHHHH HHHHHHH--- -EEEEEEE-- --EEEEEE-
M:---EHHHH-- ----HHHHHH ----HHHHHH HHHHHHHEEE -------E--

F:-HHHHHHHHH H----EEEEE E-
M:EE-------- --------EE ---------- -EEE------ HHHHHHHH-
```

(H = helix, E = strand, - = no prediction)

Figure 4.12 A 2D Structure Alignment of M.t. 0776 (M) as generated by NNPREDICT and the Ferredoxin Like Metal Binding Domain of ATX1 Metallochaperone Protein, 1CC8 (F).

4.11, the last one (with the largest E value) appears to make the most sense – especially in light of our earlier secondary structure and hydrophobicity analysis. In section 4.3.3.1, the predicted result suggests that the helices of M.t. 0776 likely surround a core of short beta strands which coalesce into a 3-4 stranded sheet. The 2D structure alignment of the two sequences is shown in table Figure 4.12. In this prediction, the α-helices are well folded and lie at the exterior of the protein, while the 3 β-sheets are buried inside. The fold identified here corresponds to a ferredoxin-like fold, suggesting that M.t. 0776 may have some metal-binding function. While these predictions can give us some good hints about the structure, we must await the results of further experiments to confirm these predictions or suggest some possible functions.

4.4 Conclusion

To summarize, the M.t. 0776 protein has been successfully expressed and purified. Likewise a protocol for the preparation of a stable NMR sample has been established. Although the structure and function of M.t. 0776 protein still remain unknown, some important characteristics of this protein can be predicted using various bioinformatics tools. These predictions suggest that M.t. 0776 has a central core of 3-4 beta strands surrounded by 3 helices in a conformation similar to that of ferredoxin-like fold. This potential structural similarity suggests that M.t. 0776 could function as a metal-binding redox protein that may be important for methane metabolism. These preliminary results could facilitate more directed

experimental studies aimed at determining the function of this protein. Further NMR studies are still needed to determine the 3D structure of this protein.

Chapter 5: General Discussion and Conclusions

This thesis has focused on the use and development of bioinformatics tools to facilitate proteomics research. Two bioinformatics tools were described in this thesis, one was developed to help in drug discovery and the other was developed to facilitate 2-DE gel archiving and comparison.

In chapter 2, I described GelScape, a web-based gel archiving system which enables users to interact with archived 2-DE gel images *via* Java applets, Javascript modules, HTML, Perl and Dynamic HTML. It was designed to enable the interaction between users and gel images and to help users analyze electronic images of 2D SDS PAGE gels. It has some of the functions that many commercial software packages have, and is freely accessible through the internet. GelScape offers five unique features or tools: 1) a hyperlinked navigation window; 2) an image-mapped gel viewer; 3) a rubber-band zooming tool; 4) a gel magnifying glass tool; and 5) a gel comparison tool. GelScape can be a useful tool for proteomics researchers and we believe that GelScape and the general concepts developed in implementing GelScape may eventually be applied to real proteomics research.

In chapter 3, I described my work on DrugBank, a tool that integrates and extracts data from many research fields, including medicinal chemistry, molecular

biology, pharmacology and structural biology. Using this tool, researchers, patients, doctors, and pharmacists may not only find direct information about a drug or drug target, but may also find related information on other web sites by clicking the hyperlinks in the DrugBank pages. DrugBank is freely accessible and platform independent, enabling people to visit it online using almost any computer with internet access. Though there are other online drug databases, DrugBank is the only one that relates drug targets with other drug information. This feature makes DrugBank a powerful tool for pharmaceutical students, professors and especially for researchers.

In chapter 4, I described my efforts aimed at puriying, expressing and characterizing the M.t.0776 protein. As shown in this chapter, it has been successfully expressed and purified. A protocol for the preparation of a stable NMR sample has been established. Additionally a 1D ^{1}H NMR trace of M.t. 0776 has been acquired. Although the structure and function of M.t.0776 are still unknown, some important characteristics of this protein can be predicted using various bioinformatics tools. It was suggested from the results of the bioinformatics analysis that M.t. 0776 has a central core of 3-4 beta strands surrounded by 3 helices in a conformation similar to that of ferredoxin-like fold. This suggests that M.t. 0776 could function in some kind of redox reactions – perhaps related to methane metabolism. The predicted results could facilitate and

rationalize more directed experimental studies aimed at determining the function and structure of this protein.

Several of the projects undertaken in this thesis are still not quite at the level where results could be published or software publicly released. I will describe some of the tasks that still remain to be done: For GelScape, in order to help users annotate their own gels, it could be useful provide several types of annotation functions in the second version of GelScape (GelScape II) which is now being developed by Nelson Young in Dr. Wishart's laboratory. To facilitate gel comparisons, a function for matching gel images through visual "registration" or image warping is also in the works. The development of a spot search or protein name utility through the implementation of a relational database system would make the archival and retrieval functions in GelScape significantly more useful. Also, it would be desirable to increase the size of the GelScape database and to integrate more protein data into the system. A CGI program could be developed to regularly update and collect data from the internet so that the GelScape database (and protein cards) could be updated automatically.

For DrugBank, we intend to significantly increase both the size and content of the DrugBank database. In addition, we intend to develop more sophisticated automated updating scripts to ensure the currency and accuracy of the data. We believe that the utility of DrugBank could be increased by including these

compounds in the database as well. In addition to expanding the database content, we also intend to extend DrugBanks search capabilities. The chemical structure search utilities could be improved with the use of Ullmann's substructure search methods (Ullmann, 1976). Likewise, the support for full proteome searches (a query proteome with 3000+ proteins) against the database still needs to be improved. The implementation of a docking utility with automated free energy evaluation (Structure Thermodynamic Calculator – Osterberg et al., 2002) is also planned. We also intend to release the source code for most of the key programs and CGI scripts used in DrugBank to encourage others to develop or improve on the code and designs. As well, the data will also be made downloadable so that others may use the information in developing more specialized databases or applications. We are hoping that this project will evolve towards becoming a "community project", allowing independent submissions, corrections, development and additions. Hopefully DrugBank may grow to be as successful and important resources as GenBank and PDB since it relates to pharmaceutical science.

For the study of M.t.0776, the predicted structure and function may rationalize and narrow the scope of future experiments, but further NMR studies are still needed to determine the 3D structure of this protein. The assignment of the backbone of this protein needs data from HNCO and ^{1}H-^{15}N TOCSY-HSQC NMR spectra. To assign the side chains for this protein, data from a HC-CH

TOCSY experiment will be needed. Other NMR experiments such as ^{1}H-^{15}N NOESY-HSQC and 2D ^{1}H-NOESY are needed to be performed to determine the full 3D structure. Once the structure of this protein is determined, it may provide a strong hint about what M.t.0776's possible function may be, through comparison with known protein structures with known functions.

As discussed in several reviews on proteomics (Patterson et al., 2003; Aebersold et al., 2003; Boguski et al., 2003; Hanash S., 2003; Wasinger, 2002), there can be little doubt that proteomics will have a significant impact on pharmaceutical and medical research in the next decade and beyond. Combined with the powerful bioinformatics tools and the ready availability of structural databases, we are expecting these methods and tools may ultimately help to lower the cost of drug discovery, to coordinate scientists' efforts worldwide and to diagnose and cure diseases that are not curable right now.

References

Abraham, D. J. and Leo, A. J. (1987) "Hydrophobicity (delta G1/2 cal)", Proteins: Structure, Function and Genetics, 2:130-152.

Aebersold, R, and Mann, M. (2003) "Mass spectrometry-based proteomics", Nature; 422:198-207

Aebersold, R. H., Leavitt, J., Saavedra, R. A., Hood, L. E., Kent, S. B. (1987) "Internal amino acid sequence analysis of proteins separated by one- or two-dimensional gel electrophoresis after in situ protease digestion on nitrocellulose"; Proc. Natl. Acad. Sci. USA, 84:6970–6974.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) "Basic local alignment search tool." J. Mol. Biol., 215:403-410.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997) "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res., 25:3389-402.

Anderson, N L.; Matheson, A. D and Steiner, S. (2000) "Proteomics: applications in basic and applied biology", Current Opinion in Biotechnology, 11:408–412

Appel, R. D., Bairoch, A., Sanchez, J. C., Vargas, J. R., Golaz, O., Pasquali, C. and Hochstrasser, D. F. (1996) "Federated two-dimensional electrophoresis database: a simple means of publishing two-dimensional electrophoresis data", Electrophoresis, 17:540-6

Bell, A.W., M.A. Ward, W.P. Blackstock, H.N.M. Freeman, J.S. Choudhary, A.P. Lewis, D. Chotai, A. Fazel, J.N. Gushue, J., Paiement, S. Palcy, E. Chevet, M. Lafreniere-Roula, R. Solari, D.Y. Thomas, A. Rowley, A. and Bergeron J.J. (2001) "Proteomics characterization of Abundant Golgi Membrane Proteins.", J Biol Chem., 276:5152-5165

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Rapp, B. A. and Wheeler, D. L. (2002) "GenBank", Nucleic Acids Res, 30:17-20

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. and Bourne, P. E. (2000) "The Protein Data Bank", Nucleic Acids Res, 28:235-242

Bishop, C. M. (1996) Neural Networks for Pattern Recognition, Oxford University Press

Bjellqvist, B., Ek, K., Righetti, P.G., Gianazza, E., Görg, A., Westermeier, R. and Postel, W. (1982) "Isoelectric focusing in immobilized pH gradients: principle, methodology and some applications", J. Biochem. Biophys. Meth., 6:317–339.

Bjellqvist, B., Hughes, G. J., Pasquali, Ch., Paquet, N., Ravier, F., Sanchez, J.-Ch., Frutiger, S. and Hochstrasser, D. F. (1993) "The focusing positions of polypeptides in immobilized pH gradients can be predicted from their amino acid sequences", Electrophoresis, 14:1023-1031.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S. and Schneider M. (2003) "The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003", Nucleic Acids Res., 31:365-370

Boguski, M. S, and McIntosh, M. W. (2003) "Biomedical informatics for proteomics", Nature; 422:233-237

Bradford, M. M. (1976) "A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding", Anal. Biochem., 72: 248

Brendel, V., Bucher, P., Nourbakhsh, I., Blaisdell, B. E. and Karlin, S. (1992) "Methods and algorithms for statistical analysis of protein sequences", Proc. Natl. Acad. Sci., USA 89: 2002-2006.

Brichory, F. M., Misek, D. E., Yim, A. M., Krause, M. C., Giordano, T. J., Beer, D. G. and Hanash, S. M. (2001) "An immune response manifested by the common occurrence of annexins I and II autoantibodies and high circulating levels of IL-6 in lung cancer", Proc Natl Acad Sci U S A., 98:9824-9829.

Briggs, M.S. and Roder, H. (1992) "Early hydrogen-bonding events in the folding reaction of ubiquitin.", Proc. Natl. Acad. Sci. USA 89: 2017-2021.

Bujard, H., Gentz, R., Lanzer, M., Stüber, D., Müller, M., Ibrahimi, I., Häuptle, M.T., and Dobberstein, B. (1987) "A T5 promotor based transcription-translation system for the analysis of proteins in vivo and in vitro", Methods Enzymol., 155:416-433.

Bull, H.B. and Breese, K. (1974) "Amino acid scale: Hydrophobicity (free energy of transfer to surface in kcal/mole)", Arch. Biochem. Biophys., 161:665-670.

Celis, J. E., Rasmussen, H. H., Gromov, P., Olsen, E., Madsen, P., Leffers, H., Honoré, B., Dejgaard, K., Vorum, H., Kristensen, D.B., Østergaard, M., Haunsø,

A., Jensen, N.A., Celis, A., Basse, B., Lauridsen, J. B., Ratz, G. P., Andersen, A.H., Walbum, E., Kjærgaard, I., Andersen, I., Puype, M., Van Damme, J. and Vandekerckhove, J. (1995) "The human keratinocyte two-dimensional gel protein database (update 1995): Mapping components of signal transduction pathways", Electrophoresis, 16: 2177-2240

Chait B.T. and Kent S.B. (1992) "Weighing naked proteins: practical, high-accuracy mass measurement of peptides and proteins.", Science, 257:1885-1894.

Christendat, D., Yee, A., Dharamsi, A., Kluger, Y., Gerstein, M., Arrowsmith, C. H., Edwards, A. M. (2000) "Structural proteomics: prospects for high throughput sample preparation", Prog Biophys Mol Biol, 73:339-345

Clauser, K. R, Hall, S. C., Smith, D. M., Webb, J. W., Andrews, L. E., Tran, H. M., Epstein, L. B. and Burlingame, A. L. (1995) "Rapid mass spectrometric peptide sequencing and mass matching for characterization of human melanoma proteins isolated by two-dimensional PAGE", Proc Natl Acad Sci U S A, 92:5072-5076

Clore, G. M. and Gronenborn, A. M. (1991) "Structures of larger proteins in solution: three- and four-dimensional heteronuclear NMR spectroscopy", Protein Science 252:1390–1399

Cooke, F. and Schofield, H. (2001) "A framework for the evaluation of chemical structure databases", J Chem Inf Comput Sci., 41:1131-1140

Daughdrill, G. W., Ackerman, J., Isern, N. G., Botuyan, M. V., Arrowsmith, C., Wold, M. S., and Lowry, D. F. (2001) "The weak interdomain coupling observed in the 70 kDa subunit of human replication protein A is unaffected by ssDNA binding", Nucleic Acids Res, 29:3270-3276

Dean, Philip M. and Zanders, Edward D. (2002) "The Use of Chemical Design Tools to Transform Proteomics Data into Drug Candidates", Computational Proteomics Supplement, 32:S28-S33.

Dongre A. R., Eng J. K. and Yates J.R. (1997) "Emerging tandem-mass-spectrometry techniques for the rapid identification of proteins", Trends Biotechnol, 15:418-425

Falquet, L., Pagni, M., Bucher, P., Hulo, N., Sigrist, C. J., Hofmann, K., and Bairoch, A. (2002) "The PROSITE database, its status in 2002.", Nucleic Acids Res., 30:235-238

Fernandez, J., Gharahdaghi, F. and Mische, S. M. (1998) "Routine identification of proteins from sodium dodecyl sulfate-polyacrylamide gel electrophoresis

(SDS-PAGE) gels or polyvinyl difluoride membranes using matrix assisted laser desorption/ionization-time of flight-mass spectrometry (MALDI-TOF-MS)", Electrophoresis, 19:1036-1045

Folmer, R. H., Geschwindner, S., and Xue, Y. (2002) "Crystal structure and NMR studies of the apo SH2 domains of ZAP-70: two bikes rather than a tandem", Biochemistry, 41:14176-14184

Fox and Edward A. (1995) "World-Wide Web and computer science reports", Communication of the ACM, 38:43-44

Frishman, D. and Argos, P. (1995) "Knowledge-based secondary structure assignment", Proteins: structure, function and genetics, 23: 566-579.

Gilbert W. (1991) "Towards a paradigm shift in biology", Nature, 349: 99

Giorgianni, F., Desiderio, D. M., and Beranova-Giorgianni, S. (2003) "Proteome analysis using isoelectric focusing in immobilized pH gradient gels followed by mass spectrometry", Electrophoresis, 24:253-259

Goate, A., Chartier-Harlin, M.-C., Mullan, M., Brown, J., Crawford, F., Fidani, L., Giuffra, L., Haynes, A., Irving, N., James, L., Mant, R., Newton, P., Rooke, K., Roques, P., Talbot, C., Pericak-Vance, M., Roses, A., Williamson, R., Rossor, M., Owen, M., and Hardy, J. (1991) "Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease", Nature 349: 704-706

Görg, A., Obermaier, C., Boguth, G., Harder, A., Scheibe, B., Wildgruber, R. and Weiss, W. (2000) "The current state of two-dimensional electrophoresis with immobilized pH gradients", Electrophoresis, 21:1037–1053.

Gottschalk A, Neubauer G, Banroques J, Mann M, Luhrmann R, and Fabrizio P. (1999) "Identification by mass spectrometry and functional analysis of novel proteins of the yeast [U4/U6-U5] tri-snRNP", EMBO Journal., 18:4535-4548

Gygi S., Han D. K., Gingras A. C., Sonenberg N., Aebersold R. (1999) "Protein analysis by mass spectrometry and sequence database searching: tools for cancer research in the post-genomic era.", Electrophoresis, 20:310-319.

Hanash, S. (2003) "Disease proteomics", Nature, 422:226-232

Hathout, Y., Riordan, K., Gehrmann, M., and Fenselau, C. (2002) "Differential protein expression in the cytosol fraction of an MCF-7 breast cancer cell line selected for resistance toward melphalan", J Proteome Res, 1:435-442

Hebebrand J, Fichter M, Gerber G, Gorg T, Hermann H, Geller F, Schafer H, Remschmidt H, and Hinney A. (2002) "Genetic predisposition to obesity in bulimia nervosa: a mutation screen of the melanocortin-4 receptor gene.", Mol Psychiatry, 7:647-651

Hecker, M., Engelmann, S. and Cordwell, S. J. (2003) "Proteomics of Staphylococcus aureus—current state and future challenges-current state and future challenges", J Chromatogr B Analyt Technol Biomed Life Sci, 787:179-195

Heinemann, U. (2000) "Structural genomics in Europe: Slow start, strong finish?" Nat. Struct. Biol. 7: 940–942

Hochstrasser, D. F., Sanchez, J.-C. and Appel, R. D (2002) "Proteomics and its trends facing nature's complexity" Proteomics, 2:807 –812

Hochuli, E. (1989) "Genetically designed affinity chromarography using a novel metal chelate absorbent", Biologically Active Molecules, 217-239

Hoogland, C., Sanchez, J.-C., Tonella, L., Binz, P.-A., Bairoch, A., Hochstrasser, D. F. and Appel, R. D. (2000) "The 1999 SWISS-2DPAGE database update", Nucleic Acids Res., 28: 286-288.

Husi, H., Ward, M. A., Choudhary, J. S., Blackstock, W. P. and Grant, S. G. (2000) "Proteomic analysis of NMDA receptor-adhesion protein signaling complexes", Nat. Neurosci., 3:661–669

James, P., Quadrom, M., Carafoli, E. and Gonnet, G. (1994) "Protein identification in DNA databases by peptide mass fingerprinting", Protein Sci., 3:1347–1350.

Janknecht, R., de Martynoff, G., Lou, J., Hipskind, R.A., Nordheim, A., and Stunenberg, H.G. (1991) "Rapid and efficient purification of native histidine-tagged protein expressed by recombinant vaccinia virus", Proc.Natl. Acad. Sci. USA 88: 8972-8976

Johnson, L. N., De Moliner, E., Brown, N. R., Song, H., Barford, D., Endicott, J. A., and Noble, M. E. (2002) "Structural studies with inhibitors of the cell cycle regulatory kinase cyclin-dependent protein kinase 2", Pharmacol Ther, 93:113-124

Jones, C. E., Holden, S., Tenaillon, L., Bhatia, U., Seuwen, K., Tranter, P., Turner, J., Kettle, R., Bouhelal, R., Charlton, S., Nirmala, N. R., Jarai, G., and Finan, P. (2003) "Expression and Characterization of a 5-oxo-6E,8Z,11Z,14Z-Eicosatetraenoic Acid Receptor Highly Expressed on

Human Eosinophils and Neutrophils.", Mol Pharmacol, 63:471-477

Kassack, M. U., Hogger, P., Gschwend, D. A., Kameyama, K., Haga, T., Graul, R. C., and Sadee, W. (2000) "Molecular modeling of G-protein coupled receptor kinase 2: docking and biochemical evaluation of inhibitors", AAPS PharmSci, 2:E2

Katoh, M., and Katoh, M. (2003) "Identification and characterization of human DAPPER1 and DAPPER2 genes in silico", Int J Oncol, 22:907-913

Kelley, L. A., Maccallum, R. and Sternberg, M. J. E. (1999) "Recognition of Remote Protein Homologies Using Three-Dimensional Information to Generate a Position Specific Scoring Matrix in the program 3D-PSSM", RECOMB 99, Proceedings of the Third Annual Conference on Computational Molecular Biology, 218-225

Kinzler, K. W. and Vogelstein, B. (1996) "Lessons from hereditary colorectal cancer" Cell, 87:159-170

Klose, J. (1975) "Protein mapping by combined isoelectric focusing and electrophoresis in mouse tissues. A novel approach to testing for induced point mutations in mammals", Humangenetik, 26: 231–243.

Krogh, A., Larsson, B., von Heijne, G. and Sonnhammer, E. L. (2001) "Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes", J Mol Biol., 305:567-580

Kwok, S, Mant, Colin T. and Hodegs, Robert S. (2002) "Importance of secondary structural specificity determinants in protein folding: Insertion of a native ß-sheet sequence into an -helical coiled-coil". Protein Science, 11:1519 –1531

Kyte, J. and Doolittle, R. F. (1982) "A simple method for displaying the hydropathic character of a protein", J Mol Biol, 157:105-132

Langen, H., Takacs, B., Evers, S., Berndt, P., Lahm, H. W., Wipf, B., Gray, C. and Fountoulakis, M. (2000) "Two-dimensional map of the proteome of Haemophilus influenzae", Electrophoresis, 21: 411-429

Lemkin, P. F. (1997) "The 2DWG meta-database of two-dimensional electrophoretic gel images on the Internet", Electrophoresis, 18:2759-2773

Lewis, T. S., Hunt, J. B., Aveline, L. D., Jonscher, K. R., Louie, D. F., Yeh, J. M., Nahreini, T. S., Resing, K. A., and Ahn, N. G. (2000) "Identification of novel MAP kinase pathway signaling targets by functional proteomics and mass

spectrometry", Mol Cell, 6:1343-1354

Lu, J. and Dahlquist, F.W. (1992) "Detection and characterization of an early folding intermediate of T4 lysozyme using pulsed hydrogen exchange and two-dimensional NMR", Biochemistry, 31:4749-4756

Lynch, S. R., Gonzalez, R. L., and Puglisi, J. D. (2003) "Comparison of X-Ray Crystal Structure of the 30S Subunit-Antibiotic Complex with NMR Structure of Decoding Site Oligonucleotide-Paromomycin Complex", Structure (Camb),11:43-53

MacGillivray, A. J. and Rickwood, D. (1974) "The heterogeneity of mouse-chromatin nonhistone proteins as evidenced by two-dimensional polyacrylamide-gel electrophoresis and ion-exchange chromatography", Eur. J. Biochem., 41:181–190.

Magga, J. M., Jarvis, S. E., Arnot, M. I., Zamponi, G. W. and Braun, J. E. A. (2000) "Cysteine string protein regulates G protein modulation of N-type calcium channels", Neuron, 28, 195–204

Maggio, Edward T. and Ramnarayan, K.; (2001); "Recent developments in computational proteomics"; Trends in Biotechnology; 19:7

Manber, U. and Bigot, P. 1998, "Connecting Diverse Web Search Facilities", Data Engineering Bulletin, 21:21-27,

Manber, U. and Wu, S. 1994, "GLIMPSE: A tool to search through entire file systems", In Proceedings of the Winter 199d USENIX Conf., 23-32. USENIX

Matouschek, A., Serrano, L., Meiering, E.M., Bycroft, M., and Fersht, A.R. (1992) "The folding of an enzyme. V. H/2H exchange-nuclear magnetic resonance studies on the folding pathway of barnase: Complementarity to and agreement with protein engineering studies", J. Mol. Biol., 224: 837-845

Matsudaira P. (1987) "Sequence from picomole quantities of proteins electroblotted onto polyvinylidene difluoride membranes", J. Biol. Chem., 262:10035–10038

McKerrow, J. H., Bhargava, V., Hansell, E., Huling, S., Kuwahara, T., Matley, M., Coussens, L., and Warren, R. (2000) "A functional proteomics screen of proteases in colorectal carcinoma", Mol Med, 6:450-460

Möller, A., Soldan, M,, Völker, U. and Maser, E. (2001) "Two-dimensional gel electrophoresis: a powerful method to elucidate cellular responses to toxic

compounds", Toxicology, 160:129-138

Morrison, P., Rosteck, P., Lai, M., Barrett, J.C., Lewis, C., Neuhausen, S., Cannon-Albright, L., Goldgar, D., Wiseman, R., Kamb, A. and Skolnick, M. H. (1994) "A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1."; Science, 266:6671

Mullins, L.S., Pace, C.N., and Raushel, F.M. (1993) "Investigation of ribonuclease-tl folding intermediates by hydrogen-deuterium amide exchange-2D-NMR spectroscopy", Biochemistry, 32:6152-6156

Nakai, K. and Horton, P. (1999) "PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization", Trends Biochem Sci., 24:34-36.

Nauli, S., Kuhlman, B., Le Trong, I., Stenkamp, R. E., Teller, D., and Baker, D. (2002) "Crystal structures and increased stabilization of the protein G variants with switched folding pathways NuG1 and NuG2", Protein Sci, 11:2924-2931

Neubauer G., King A., Rappsilber J., Calvio C., Watson M. and Ajuh P. (1998) "Mass spectrometry and EST-database searching allows characterization of the multi-protein spliceosome complex", Nat Genet, 20:46-50.

Norwell, J.C. and Machalek, A.Z. (2000) "Structural genomics programs at the US National Institute of General Medical Sciences", Nat. Struct. Bio., 7, 931

O'Farrell, P. H. (1975) "High resolution two-dimensional electrophoresis of proteins", J. Biol. Chem., 250: 4007–4021

Orengo, C. A., Brown, N. P. and Taylor, W. R. (1992) "Fast structure alignment for protein databank searching", Proteins, 14:139-167

Osterberg, F., Morris, G. M., Sanner, M. F., Olson, A. J. and Goodsell, D. S. (2002) "Automated docking to multiple target structures: incorporation of protein mobility and structural water heterogeneity in AutoDock", Proteins, 46:34-40

Otyepka, M., Krystof, V., Havlicek, L., Siglerova, V., Strnad, M., and Koca, J. (2000) "Docking-based development of purine-like inhibitors of cyclin-dependent kinase-2", J Med Chem, 43: 2506-2513

Pandey, A., and Mann, M. (2000) "Proteomics to study genes and genomes", Nature, 405: 837 –846

Patterson, S. D., and Aebersold, R. H. (2003) "Proteomics: the first decade and beyond", Nat Genet, 33 Suppl:311-323

Patton, W. F. (2002) "Detection technologies in proteome analysis", J Chromatogr B Analyt Technol Biomed Life Sci, 771:3-31

Pennacchio, L. A. and Rubin, E. M. (2003) "Comparative genomic tools and databases: providing insights into the human genome", J Clin Invest, 111:1099-1106

Perkins, D. N., Pappin, D. J., Creasy, D. M. and Cottrell, J. S. (1999) "Probability-based protein identification by searching sequence databases using mass spectrometry data", Electrophoresis, 20:3551-3567

Pleissner, K.-P., Eifert, T. and Jungblut, P.R. (2002) "A European pathogenic microorganism proteome database:Construction and maintenance", Comparative and Functional Genomics (Comp.Funct.Genom.), 3: 97-100.

Plomin, R., McClearn, G. E., Smith, D. L., Vignetti, S., Chorney, M. J., Chorney, K., Venditti, C. P., Kasarda, S., Thompson, L. A., and Detterman, D. K. (1994) "DNA markers associated with high versus low IQ: the IQ quantitative trait loci (QTL) project", Behav. Genet., 24:107-118

Porath, J., Carlsson, J., Olsson, I., and Belfrage, G. (1975) "Metal chelate affinity chromatography, a new approach to protein fractionation", Nature, 258:598-599

Perczel, A., Jakli, I., McAllister, M. A. and Csizmadia, I. G. (2003) "Relative stability of major types of beta-turns as a function of amino acid composition: a study based on Ab initio energetic and natural abundance data", Chemistry, 9:2551-2566.

QIAGEN (2001) "The QIAexpressionist Fifth Edition", 18-23

Quadroni, M. and James, P. (1999) "Proteomics and automation", Electrophoresis, 20:664-677

Rabilloud, T., Kieffer, S., Procaccio, V., Louwagie, M., Courchesne, P. L., Patterson, S. D., Martinez, P., Garin, J. and Lunardi, J. (1998) "Two dimensional electrophoresis of human placental mitochondria and protein identification by mass spectrometry. Toward a human mitochondrial proteome", Electrophoresis, 19: 1006-1014.

Rabilloud, T., Vuillard, L., Gilly, C., Lawrence, J. J. (1994) "Silver-staining of proteins in polyacrylamide gels: a general overview", Cell Mol Biol (Noisy-le-grand), 40:57-75

Radford, S. E., Buck, M., Topping, K. D., Dobson, C. M., and Evans, P. A. (1992) "Hydrogen exchange in native and denatured states of hen egg-white lysozyme", Proteins Struct. Funct. Genet., 14:237-248.

Rebhan, M., Chalifa-Caspi, V., Prilusky, J. and Lancet, D. (1997) "GeneCards: integrating information about genes, proteins and diseases", Trends Genet, 13:163

Rost, B. (2001) "Protein secondary structure prediction continues to rise", Journal of Structural Biology, 134:204-218

Sagliocco, F., Guillemot, J.-C., Monribot, C., Capdevielle, J., Perrot, M., Ferran, E., Ferrara, P. and Boucherie, H. (1996) "Identification of Proteins of the Yeast Protein Map using Genetically Manipulated Strains and Peptide-Mass Fingerprinting", Yeast 12:1519-1534.

Sali, A. (1998) "100,000 protein structures for the biologist", Nat Struct Biol., 5:1029-1032

Sali, A., Glaeser, R., Earnest, T., and Baumeister, W. (2003) "From words to literature in structural proteomics", Nature, 422:216-225

Schagger, H., Aquila, H., Von Jagow, G. (1988) "Coomassie blue-sodium dodecyl sulfate-polyacrylamide gel electrophoresis for direct visualization of polypeptides during electrophoresis", Anal Biochem, 173:201-205

Schagger, H. and von Jagow, G. (1987) "Tricine-sodium dodecyl sulfate-polyacrylamide gel electrophoresis for the separation of proteins in the range from 1 to 100 kDa", Anal. Biochem., 166:368-379

Schwartz, E., Scherrer, G., Hobom, G., and Hössel, H. (1978) "Nucleotide sequence of cro, cll and part of the 0 gene in phage lambda DNA", Nature, 272: 410-414

Serrano, L., Sancho, J., Hirshberg, M.and Fersht, A R. (1992) "Alpha-helix stability in proteins. I. Empirical correlations concerning substitution of side-chains at the N and C-caps and the replacement of alanine by glycine or serine at solvent-exposed surfaces", J Mol Biol., 227:544-559

Shevchenko, A., Jensen, O. N., Podtelejnikov A. V., Sagliocco, F., Wilm, M., Vorm, O., Mortensen, P., Shevchenko, A., Boucherie, H., and Mann, M. (1996) "Linking genome and proteome by mass spectrometry: Large-scale identification of yeast proteins from two dimensional gels", Proc. Natl Acad. Sci. USA, 93:14440-14445

Shevchenko, A., Wilm, M., Vorm, O. and Mann, M (1996) "Mass spectrometric

sequencing of proteins from silver stained polyacrylamide gels", Anal. Chem., 68:850-858

Smith, D. R., Doucette-Stamm, L. A., Deloughery, C., Lee, H.-M., Dubois, J., Aldredge, T., Bashirzadeh, R., Blakely, D., Cook, R., Gilbert, K., Harrison, D., Hoang, L., Keagle, P., Lumm, W., Pothier, B., Qiu, D., Spadafora, R., Vicare, R., Wang, Y., Wierzbowski, J., Gibson, R., Jiwani, N., Caruso, A., Bush, D., Safer, H., Patwell, D., Prabhakar, S., McDougall, S., Shimer, G., Goyal, A., Pietrovski, S., Church,G. M., Daniels, C. J., Mao, J.-i., Rice, P., Nolling, J. and Reeve, J. N. (1997) "Complete genome sequence of Methanobacterium thermoautotrophicum deltaH: functional analysis and comparative genomics", J. Bacteriol., 179:7135-7155

Srinivas, P. R., Srivastava, S., Hanash, S., and Wright, G. L. (2001) "Proteomics in Early Detection of Cancer", Clinical Chemistry, 47:1901-1911

Srinivas, P. R., Verma, M., Zhao, Y. and Srivastava, S. (2002) "Proteomics for Cancer Biomarker Discovery", Clinical Chemistry, 48:1160-1169

Steiner, S., Aicher, L., Raymackers, J., Meheus, L., Esquer-Blasco, R., Anderson, L., Cordier, A.,(1996) "Cyclosporine A mediated decrease in the rat renal calcium binding protein calbindin-D 28 kDa", Biochem. Pharmacol., 51:253–258.

Stüber, D., Matile, H., and Garotta, G. (1990) "System for high-level production in Escherichia coli and rapid purification of recombinant proteins: application to epitope mapping, preparation of antibodies, and structure-function analysis." In: Immunological Methods, Lefkovits, I. and Pernis, B., eds., vol. IV, Academic Press, New York, pp.121-152.

Sulkowski, E. (1985) "Purification of proteins by IMAC". Trends Biotechnol., 3:1-7

Sundararaj, S., Habibi-Nazhad, B., Rouani, M., Guo, A., Young, S., Ellison, M., and Wishart, D. S., "The CyberCell Database: A Comprehensive, Self-updating Relational Database for Modeling E. coli", *Nucleic Acids Research*, submitted

Sutcliffe, J. G. (1979) "Complete nucleotide sequence of the E. Coli plasmid pBR322", Cold Spring Harbor Symp. Quant. Biol. 43:77-90.

Taylor, W. R. and Orengo, C. A. (1989) "Protein structure alignment", J. Mol.Biol. 208:1-22.

Terwilliger, T. C. (2000) "Structural genomics in North America", Nat. Struct. Biol., 7:935–939

The Genome International Sequencing Consortium (2001) "Initial sequencing and analysis of the human genome", Nature, 409:860–921

Thompson, L. K., McDermott, A. E., Raap, J., van der Wielen, C. M., Lugtenburg, J., Herzfeld, J. and Griffin, R. G. (1992) "Rotational resonance NMR study of the active site structure in bacteriorhodopsin: conformation of the Schiff base linkage", Biochemistry, 31:7931-7938

Tobias, D. J., Sneddon, S. F. and Brooks, C. L. 3rd (1992) "Stability of a model beta-sheet in water", J. Phys. Chem., 96:3864

Uhrig, J. F., Soellick, T. R., Minke, C. J., Philipp, C., Kellmann, J. W. and Schreier, P. H. (1999) "Homotypic interaction and multimerization of nucleocapsid protein of tomato spotted wilt tospovirus: Identification and characterization of two interacting domains", Proceedings of the National Academy of Sciences of the United States of America., 96:55-60

Ullmann, J. R. (1976) "An algorithm for subgraph isomorphism", J. ACM, 23:31-42

Vankayalapati, H., Bearss D. J., Saldanha, J. W., Munoz, R. M., Rojanala, S., Von Hoff, D. D., and Mahadevan D. (2003) "Targeting aurora2 kinase in oncogenesis: a structural bioinformatics approach to target validation and rational drug design", Mol Cancer Ther, 2:283-294

Varley, P., Gronenborn, A. M., Christensen, H., Wingfield, P. T., Pain, R. H., and Clore, G. M. (1993) "Kinetics of the folding of an all sheet protein interleukin-1 beta", Science, 260:1110-1113

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A.G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern,

D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead. M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigo, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A. and Zhu, X. (2001) "The sequence of the human genome", Science, 291:1304–1351.

Wall, L., Christiansen, T. and Schwartz, R. L. (1996) Programming Perl, 2nd edn. O'Reilly & Associates, Inc., Sebastopol, CA. US

Walters, P. and Stahl, M. (1992) Dolata Research Group Department of Chemistry, University of Arizona, Tucson, AZ 85721, babel@mercury.aichem.arizona.edu

Wang, G. (2002) "How the lipid-free structure of the N-terminal truncated human apoA-I converts to the lipid-bound form: new insights from NMR and X-ray structural comparison", FEBS Lett, 529:157-161

Wang, H. C., Xia, Y. and He, Y. (1996) "The Rapid Identification of Protein Through Its Amino Acid Composition", Sheng Wu Hua Xue Yu Sheng Wu Wu Li Xue Bao (Shanghai), 28:293-299

Wasinger, V. C. and Corthals, G. L. (2002) "Proteomic tools for biomedicine", Journal of Chromatography B, 771:33-48

Weininger, D. (1988) "SMILES, a Chemical Language and Information System," J. Chem. Info. Comp. Sci., 28, 31

Wishart, D. S. and Sykes, B. D. (1994) "Chemical shifts as a tool for structure determination", Methods Enzymol., 239:363-392

Wishart, D. S. (2000) "Course Notes from PHARM 580, Lecture 2", Faculty of Pharmaceutical Science, University of Alberta

Wüthrich, K. (1986) NMR of Proteins and Nucleic Acids. New York: John Wiley
Yanagida, Mitsuaki; (2002); "Functional proteomics; current achievements"; Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences; 771:89-106

Yates J, 3rd. (1998) "Mass spectrometry and the age of the proteome." J Mass Spectrom, 33:1-19.

Yates, J. R., Speicher, S., Griffin, P. R. and Hunkapiller, T. (1993) "Peptide mass maps: a highly informative approach to protein identification", Anal. Biochem., 214:397–408.

Yee, A., Pardee, K., Christendat, D., Savchenko, A., Edwards, A. M., and Arrowsmith, C. H. (2003) "Structural proteomics: toward high-throughput structural biology as a tool in functional genomics", Acc Chem Res, 36:183-189

Yokoyama, S., Matsuo, Y., Hirota, H., Kigawa, T., Shirouzu, M., Kuroda, Y., Kurumizaka, H., Kawaguchi, S., Ito, Y., Shibata, T., Kainosho, M., Nishimura, Y., Inoue, Y., and Kuramitsu S. (2000) "Structural genomics projects in Japan". Nat. Struct. Biol. 7:943–945

Zarembinski, T. I., Hung, L. W., Mueller-Dieckmann, H. J., Kim, K. K., Yokota, H., Kim, R. and Kim, S. H. (1998) "Structure-based assignment of the biochemical function of a hypothetical protein: a test case of structural genomics", Proc Natl Acad Sci U S A., 95:15189-193

Zeikus, J. G., and Wolfe, R. S. (1972) "Methanobacterium thermoautotrophicus sp. n., an anaerobic, autotrophic, extreme thermophile", J. Bacteriol., 109:707-713